DOCUMENT RESUME

ED 153 095 CG 012 338

AUTHOR	Schneider, Benjamin, Ed.
TITLE	Some Conceptual and Methodological Issues in Understanding Ability-Performance Relationships. Research Report No. 16.
INSTITUTION	Maryland Univ., College Park. Dept. of Psychology.
SPONS AGENCY	Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE	Aug 77
CONTRACT	N00014-75-C-0884
NOTE	246p.
EDRS PRICE	MF-$0.83 HC-$12.71 Plus Postage.
DESCRIPTORS	*Ability Identification; Conference Reports; Content Analysis; Methods; *Motivation Techniques; *Performance Factors; *Personnel Selection; *Predictor Variables; Sex Stereotypes; *Test Bias

ABSTRACT

This technical report contains the text of papers presented at a conference on personnel selection research. The conference brought together scholars to share their ideas regarding ways in which their own special interests help shed light on ability-performance relationships observed in personnel selection studies. The contributions can be divided into four sets, based on their focus: methodological, cautionary, organization entry, and organization process. Two major implications are presented: (1) the chances of finding statistically significant algebraic interactions of ability and other variables (race, sex, situation, motivation) are poor; and (2) in addition to more traditional methodological reasons (inadequate measurement), at a conceptual level algebraic interaction terms will not generally be found to be statistically significant in field studies because of naturally occurring interactions among people. The conclusion is reached that ability and work condition, in linear combination, should be used by personnel selection researchers to increase both the prediction of performance and the understanding of work behavior. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED153095

SOME CONCEPTUAL AND METHODOLOGICAL ISSUES IN UNDERSTANDING ABILITY-PERFORMANCE RELATIONSHIPS

BENJAMIN SCHNEIDER, EDITOR

CONTRIBUTIONS BY:

C. J. BARTLETT
PHILIP BOBKO
ROBERT M. GUION
ROBERT L. [illegible]NAN
JOHN E. HUNTER
BRUCE L. KATCHER
EDWIN A. LOCKE
ANTHONY J. MENTO
STEVEN B. MOSIER
WILLIAM A. OWENS
VIRGINIA E. SCHEIN
FRANK L. SCHMIDT
BENJAMIN SCHNEIDER
JOHN P. WANOUS

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Research Report No. 16

August, 1977

The papers contained in this Technical Report were presented at a Personnel Selection Research Conference sponsored by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research under Contract No. N00014-75-C-0884, Contract Authority Identification Number, NR 151-375, Benjamin Schneider and C. J. Bartlett, Principal Investigators.

Reproduction in whole or part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited.

psychology

UNIVERSITY OF MARYLAND
1807-1856-1920

CG012338

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report No. 16	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Some Conceptual and Methodological Issues in Understanding Ability-Performance Relationships		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Benjamin Schneider, Editor		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0884
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Maryland College Park, Maryland 20742		10. PROGRAM ELEMENT PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N RR 042-04; RR 042-04-02 NR 151-375
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research (Code 458) Arlington, Virginia 22217		12. REPORT DATE August, 1977 13. NUMBER OF PAGES iv. + 232
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Personnel Selection; Ability-Performance Relationship; Content Validity; Moderator Variables; Test Bias; Differential Validity; Realistic Job Preview; Sex Role Stereotyping; Ability x Motivation Interaction; Person x Situation Interaction		

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This Technical Report contains the text of invited papers presented at the Personnel Selection Research Conference, Department of Psychology, University of Maryland, March 3-4, 1977. The conference brought together scholars to share their ideas regarding ways in which their own special interests help understand ability-performance relationships observed in personnel selection studies.

Robert M. Guion (Bowling Green State University) presented "Content Validity in Moderation: Cautions Concerning Fairness." This paper addresses some of the

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

problems with, and implications of, the uncritical acceptance of the notion of content validity.

Frank L. Schmidt (Civil Service Commission) presented "Moderator Research and the Law of Small Numbers" (co-authored with John E. Hunter of Michigan State University). This paper proposes, and presents some data supporting, a model of validity generalization that seriously questions the extent to which one may expect moderators (such as race and situation) to have an impact on validity coefficients.

C. J. Bartlett (University of Maryland, College Park) presented "Testing for Fairness with a Moderated Multiple Regression Strategy: An Alternative to Differential Analysis" (co-authored with Philip Bobko, Steven B. Mosier, and Robert L. Hannan, University of Maryland, College Park). These authors argue that an adequate examination of test fairness requires an analysis of differential prediction involving slopes and intercepts of regression lines best accomplished through the application of a moderated multiple regression strategy.

William A. Owens (The University of Georgia) presented "Moderators and Subgroups." This presentation concentrates on subgrouping of subjects and using subgroup membership as the approach of choice for enhancing meaning and prediction.

John P. Wanous (Michigan State University) presented "Realistic Job Previews: Can a Procedure to Reduce Turnover Also Influence the Relationship Between Abilities and Performance?" This paper suggests a somewhat pessimistic view of the potential for realistic job previews to impact ability-performance relationships.

Virginia E. Schein (University of Pennsylvania) presented "The Effects of Sex Role Stereotyping on the Ability-Performance Relationship: Prior Research and New Directions." The lack of research on the effects of sex role stereotyping on the performance of women in management is noted and it is suggested that the way in which sex role stereotypical thinking limits women's ability to acquire power may impact ability-performance relationships.

Edwin A. Locke (University of Maryland, College Park) presented "The Interaction of Ability and Motivation in Performance: An Exploration of the Meaning of Moderators" (co-authored with Anthony J. Mento and Bruce L. Katcher, University of Maryland, College Park). An hypothesis is presented and tested that one explanation of moderator effects is that they are due to different degrees of homogeneity with respect to a causal variable among different subgroups. It was found that ability predicted performance better in groups which were homogeneous with respect to motivation that in those which were motivationally heterogeneous.

Benjamin Schneider (University of Maryland, College Park) presented "Person-Situation Selection: A Review of Some Ability-Situation Interaction Research." This review suggests that when reward system, job, and work climate reward, support, and encourage the display of ability then validity for ability measures, and performance levels, will both be high.

Finally, Schneider presents a brief overview of the various contributions deriving some implications from these papers for integrating the more psychometric and organizationally-oriented approaches to the prediction and understanding of employee behavior.

Abstract

This technical report contains the text of invited papers presented at the Personnel Selection Research Conference, Department of Psychology, University of Maryland, March 3-4, 1977. The conference brought together scholars to share their ideas regarding ways in which their own special interests help understand ability-performance relationships observed in personnel selection studies.

Robert M. Guion (Bowling Green State University) presented "Content Validity in Moderation: Cautions Concerning Fairness." This paper addresses some of the problems with, and implications of, the uncritical acceptance of the notion of content validity.

Frank L. Schmidt (Civil Service Commission) presented "Moderator Research and the Law of Small Numbers" (co-authored with John E. Hunter of Michigan State University). This paper proposes, and presents some data supporting, a model of validity generalization that seriously questions the extent to which one may expect moderators (such as race and situation) to have an impact on validity coefficients.

C. J. Bartlett (University of Maryland, College Park) presented "Testing for Fairness with a Moderated Multiple Regression Strategy: An Alternative to Differential Analysis" (co-authored with Philip Bobko, Steven B. Mosier, and Robert L. Hannan, University of Maryland,

College Park). These authors argue that an adequate examination of test fairness requires an analysis of differential prediction involving slopes and intercepts of regression lines best accomplished through the application of a moderated multiple regression strategy.

William A. Owens (The University of Georgia) presented "Moderators and Subgroups." This presentation concentrates on subgrouping of subjects and using subgroup membership as the approach of choice for enhancing meaning and prediction.

John P. Wanous (Michigan State University) presented "Realistic Job Previews: Can a Procedure to Reduce Turnover Also Influence the Relationship Between Abilities and Performance?" This paper suggests a somewhat pessimistic view of the potential for realistic job previews to impact ability-performance relationships.

Virginia E. Schein (University of Pennsylvania) presented "The Effects of Sex Role Stereotyping on the Ability-Performance Relationship: Prior Research and New Directions." The lack of research on the effects of sex role stereotyping on the performance of women in management is noted and it is suggested that the way in which sex role stereotypical thinking limits women's ability to acquire power may impact ability-performance relationships.

Edwin A. Locke (University of Maryland, College Park) presented "The Interaction of Ability and Motivation in Performance: An Exploration of the Meaning of Moderators" (co-authored with Anthony J. Mento and Bruce L. Katcher, University of Maryland, College Park).

An hypothesis is presented and tested that one explanation of moderator effects is that they are due to different degrees of homogeneity with respect to a causal variable among different subgroups. It was found that ability predicted performance better in groups which were homogeneous with respect to motivation than in those which were motivationally heterogeneous.

Benjamin Schneider (University of Maryland, College Park) presented "Person-Situation Selection: A Review of Some Ability-Situation Interaction Research." This review suggests that when reward system, job, and work climate reward, support, and encourage the display of ability then validity for ability measures, and performance levels, will both be high.

Finally, Schneider presents a brief overview of the various contributions deriving some implications from these papers for integrating the more psychometric and organizationally-oriented approaches to the prediction and understanding of employee behavior.

Contents

Content Validity in Moderation: Cautions Concerning Fairness 1
Robert M. Guion

Moderator Research and the Law of Small Numbers 31
Frank L. Schmidt and John E. Hunter

Testing for Fairness with a Moderated Multiple Regression Strategy: An Alternative to Differential Analysis 75
C. J. Bartlett, Philip Bobko, Steven B. Mosier, and Robert L. Hannan

Moderators and Subgroups 95
William A. Owens

Realistic Job Previews: Can a Procedure to Reduce Turnover Also Influence the Relationship between Abilities and Performance 115
John P. Wanous

The Effects of Sex Role Stereotyping on the Ability-Performance Relationship: Prior Research and New Directions 133
Virginia E. Schein

The Interaction of Ability and Motivation in Performance: An Exploration of the Meaning of Moderators 157
Edwin A. Locke, Anthony J. Mento, and Bruce L. Katcher

Person-Situation Selection: A Review of Some Ability-Situation Interaction Research 183
Benjamin Schneider

Summary and Implications of the Conference: A Personal View 217
Benjamin Schneider

Content Validity in Moderation: Cautions Concerning Fairness

Robert M. Guion

Bowling Green State University

Abstract

"Content validity" has been widely but unwisely hailed as a solution to many problems in employee selection. The enthusiasm and its scope must be tempered. The arguments of this paper begin with the recognition that sampling from content domains cannot logically be substituted for criterion-related validity. Moreover, it must be recognized that the psychometric notion of validity ordinarily refers to scores, not to the stimulus content of a test. It is therefore suggested that evaluations of scores be based on the principle of construct validation; that is, that possible sources of contamination be considered as alternative hypotheses about what is measured by a content sample as it is scored.

Content Validity in Moderation: Cautions Concerning Fairness

Robert M. Guion
Bowling Green State University

According to erroneous hearsay, the ostrich sticks its head in the sand in the face of possible danger. Testing specialists have played ostrich--replacing sand with the notion of content validity. The notion of content validity is, to change metaphors, too often seen as a panacea for all technical and litigious problems now surrounding the use of criterion-related validity.

There is much to be uneasy about in the trend toward carefree acceptance of this notion; in a previous paper, among other problems, I identified the "untouched problem of fairness" as one source of discontent (Guion, 1977). In this paper, I should like to touch on what is sometimes called the "content validity approach to validation" in the context of a discussion of "the ability-performance relationship."

The Ability-Performance Relationship

In a classic understanding of the ability-performance relationship, one has on the left-hand side of a mathematically-defined

functional equation a Y variable, called a criterion, which is, directly or at some stage removed, a reflection if not a measure of performance, and, on the right-hand, an X variable, called a predictor, which is, perhaps also at some stage removed, a reflection if not a measure of ability. That sentence is also a classic of sorts--a classic example of obfuscating waffling. In statistical language, it can be clarified simply by saying that the relationship is the regression of a performance variable on a measure of ability.

However, the waffling seems necessary. Psychometric research has not yielded clear prescriptions for either of the variables, and the variety of potential regression forms is great. Consider performance: Performance is measured by nearly any convenient means, usually ratings, and these means are often many stages removed from the actual performance itself. One of the illustrative experiences (and one that has done most to destroy the optimistic innocence of my youth) has been the discovery that many police sergeants seem to rate police performance in terms of appearance. I ask, "How is he at fighting crime?" and the answer may well be a variant of, "He must be good because he <u>looks</u> like a good cop." Ratings, even production records, may be used as performance measures, but the tired old discussions of the "criterion problem" must always take note that the actual behavior or "performance" is often inferred from a great distance.

Another reason for waffling is that the term ability is itself exceedingly ambiguous. Munsterberg said, "But the developing of abilities does not refer only to external acts like reading and writing, but just as much to intellectual activities like attending, thinking, calculating" (Munsterberg, 1915, p. 375). A few years later Spearman (1927) wrote an entire book entitled "The Abilities of Man"; it was a book on intelligence. Hull (1928) appeared to distinguish abilities from aptitudes, the distinction apparently being that abilities are more complex attributes than aptitudes. More recent thinking, as nearly as I can detect it from the books on my shelves, tends to treat abilities in categories such as intellectual or motor or social abilities or skills--that is, in terms of very broad attributes or traits relevant to a variety of settings. Abundance in such attributes may be essential prerequisites to abilities defined with reference to jobs; the latter is somehow inferred from the former.

There are many contexts in which the term ability has this narrower focus defined by a job or task. That is, we may speak of typing ability, or of the ability to drive a car, or the ability to handle some aspect of a job. In these cases, we speak not of aptitudes to learn how to do these things but of the level of current skill in doing them. In this sense, the term ability focuses less on the individual traits than on the task to be performed.

When this narrow focus is what we mean by ability, the phrase

"ability-performance relationship" is redundant to whatever extent ability is defined by performance.

Evaluations of Measures of Ability

With either meaning of the word, a measure of an ability must be evaluated if it is to be used as a basis for decision; that evaluation ordinarily culminates in a judgment concerning the validity of the measurements taken. If ability means aptitude, then the evaluation of the ability measure generally requires the computation of a correlation coefficient as a statement of predictive (or more generally, criterion-related) validity. Aptitude is not a word that stands alone; when one speaks of aptitude, one refers to an aptitude for some sort of subsequent performance. The measure of performance is, then, in Hull's (1928) terms, the "actual aptitude"; it is the criterion with which the aptitude measure is to be correlated in what we now call criterion-related validity. If the correlation is quite high, or if it is at least statistically significant, the aptitude measure is declared valid. The criterion-related validation approach to validity has been considered straightforward: one gives the aptitude measure to a bunch of people called a sample, criterion measures are obtained on the same bunch of people (preferably at some later time), and the two sets of measurements are correlated.

It is worth mentioning in this general oversimplification that aptitude measures of ability may also be evaluated in terms of con-

struct validity. The rules and procedures are more complex and more uncertain, but the conclusion, when reached, may be more structured and more generalizable. Unfortunately, even if one concludes that the measure is valid for a certain construct, one may still be left with the uncertainty of whether the validly measured construct is at all related to performance. That is, in establishing the construct validity of the aptitude measure as a measure, the validity of the hypothesis of a relationship between ability and performance remains untested.

This leads to my first main point: where ability means aptitude, (a) the relationship of that ability to performance on a given job is a hypothesis, and (b) criterion-related validation is intended to show the validity, not of the test, but of that hypothesis (Guion, 1976).

Since nearly all employment testing implies such a hypothesis, at least in part, the underlying logic of criterion-related validation is pervasive. Nevertheless, there are many problems associated with it. One of these is the fact that the ability-performance relationship may differ in samples from identifiably different populations. Another is the practical fact that sample sizes are almost always too small, especially for any subgrouping. Another is that damnable criterion problem. Another is that sample statistics may differ grossly from population parameters; in a monte carlo simulation done at Bowling Green, fifty samples of 200 cases each were not quite suf-

ficient for a stable sampling distribution (Jones, Note 1). This may be due to the use of error-free data in the simulation; it's rather embarrassing, however, to depend on measurement error to level out sampling error!

Such problems literally demand an alternative to the traditional requirement of independent criterion-related validation research in each new situation ("where technically feasible"), and many people have fallen all over themselves in their hurry to proclaim something called content validity a savior for employment testing. It certainly does seem relevant. Even a change in language proclaims its relevance. In criterion-related validity, the hypothesis is that $Y = f(X)$, where Y and X are clearly different things. With so-called content validity, however, the language is changed to read that X is a sample, not merely a function, of Y; i.e., X and Y are the same things, differing only in amount. In criterion-related validation, we evaluate the measure of ability in terms of the strength of its functional relationship with the measure of performance; in content validity, we evaluate the measure of ability in terms of how faithfully it represents the performance.

So this is the second main point: when one refers to content validity, one is not looking at an "ability-performance relationship" at all; one is looking at ability as a part of performance. Content sampling implies at least partial identity or overlap, not a functional relationship between measures of different things. Therefore,

the kinds of concerns over fairness to subgroups that are so very important in discussions of criterion-related validity are simply irrelevant to valid content samples.

Unfortunately, this is too glib; there are other problems of fairness to consider.

The Ephemeral Notion of Content Validity

Equally unfortunately, glibness has characterized this notion of content validity since its intrusion into employment testing in the original OFCC Testing Order (OFCC, 1968). It may be useful to review the process of disillusionment about content validity. In Standards for Educational and Psychological Tests, we said, "An investigation of content validity requires that the test developer or test user specify his objectives and carefully define the performance domain in light of those objectives" (APA, 1974, p. 28). And, "Content validity is determined by a set of operations, and one evaluates content validity by the thoroughness and care with which these operations have been conducted" (p. 29). The Standards, unfortunately, are not informative about either the definition of the performance domain to be sampled or the operations of such sampling. The Standards are mute on questions of scoring a valid content sample.

The Principles for the Validation and Use of Personnel Selection Procedures (Division of Industrial-Organizational Psychology, 1975) is perhaps a bit clearer. It says (a) that a test developer should

define a job content domain, (b) that the definition should be given in terms of tasks, activities, or responsibilities, or perhaps job knowledge, (c) that the sample should include all important aspects of the domain, and (d) that the qualifications of people who make various kinds of judgments in the process should be duly recorded. In short, the Principles speak less of evaluating an existing instrument than of the procedure for developing an appropriate content sample.

I have been profoundly dissatisfied with these two statements. The material on content validity in the Standards is filled with contradiction and confusion; it illustrates the fallacy of using democratic machinery to resolve a technical issue. The Principles also represent a bow to the opinions of committee members whose attention to the matter had been casual. Mary Tenopyr and I, considering such materials as Messick's Presidential address to the Division on Measurement and Evaluation (Messick, 1975) and Ebel's commentary at Content Validity II (Ebel, 1977), among others, had reached the conclusion others have reached--that there is really no such thing as content validity, that there is only content-oriented test construction (Tenopyr, 1977).

My discontent with the notion of content validity was strong, and I heartily wished the whole notion would go away. In fact, the paper on content validity for the Division on Measurement and Evaluation (Guion, 1977) took the nursery rhyme for its theme:

Last night I saw upon the stair
A little man who wasn't there.
He wasn't there again today.
Oh, how I wish he'd go away!

Too many people, on committees or off, thought they had something great in the notion of content validity, and the notion was not going to go away merely because some of us considered it an ephemeral, erroneous notion. So I tried to examine what people meant when they used the term. My conclusions: ". . . people who talk about content validity are either (a) talking about a special case of construct validity or (b) not talking about validity at all, but simply about the operational definitions of their constructs . . . . people who talk about content validity are talking about important evaluations of operational definitions. They invoke the concept of content validity primarily justifying the use of a measuring instrument" (Guion, 1977, p. 5).

It was necessary, therefore, to explore the requirements of a justified instrument. Five were named. None of these five had anything at all to do with the technical issue of the scoring of the content sample. Here, I submit, is the vulnerability of discussions of content validity in the justification of test use.

Last fall, at the conference of the Personnel Testing Council of Southern California known as Content Validity II.5, I finally came to the measurement issues involved. Before discussion of them, let me

now make my third main point: Although content-oriented test development may assure a representative sample of the tasks, activities, or responsibilities in the job content domain, or of a job knowledge domain, the evaluation of the score assigned to performance on that sample must be evaluated according to the principle of construct validity. That principle is the notion of alternative interpretations of scores.

Tenopyr (1977) has said it first: ". . . there should be no real conflict about whether content or construct validation is appropriate in a given situation. The question instead is one of for which class of constructs is evidence of traditional views of content validity alone enough to justify the contention that these constructs are being measured." Traditional view here refers simply to adequacy of sampling. What she has recognized plainly is that ability, whether in the braod sense of an aptitude, or in the narrow sense of a job skill, is indeed a construct, and the measurement of a construct is subject to contaminating sources of error.

Let us be very clear about this before going on. Validity, in any form, is a statement about the evaluation of the use of a measure. It is not an autonomous attribute of the test; it is an attribute of the inferences to be derived from scores. We do not validate tests; we validate inferences. If we can draw any inference at all from what might be called a valid content sample (i.e., a representative one), it is that performance in the domain sampled will be

reflected in performance on the sample itself. The performance itself is only rarely quantitative; we turn a sample of behavioral content into a measure of performance when we establish a set of operations--operations which are not in the job content domain itself--for assigning numbers to the performance. And when we do this, the numbers are intended to represent the ability basic to the observed performance. That ability is a construct, and the numbers used to represent that construct must, if they are valid, fit into a logical, or nomological, network of relationships.

In short, while we are constructing an ability test by content sampling, we may speak of the validity of the sample and mean by that its representativeness of the defined job content domain. When we devise a scheme for quantifying quality of performance on that sample, however, we have an obligation to be sure that the number system validly reflects the underlying ability. We cannot simply assume, because we have a representative sample of content, that we also have a valid number system. And since the psychometric concept of validity refers, not to the stimulus material or accompanying constraints on responses, but to the inferences from the number system, there simply is no such thing as content validity of scores--that is, of valid inferences from scores based solely on test content. Let that be my fourth main point.

For those who may feel that the denial of content validity is a heresy for which there is no salvation, I would abstract a bit here

from the earlier paper in which, in line with Ebel (1961), I argued that one does not need to invoke a concept of validity in evaluating every test. In that paper (Guion, 1977) I suggested five requirements which, if met, would justify the use of a test as an operational definition of a variable, without any empirical investigations into validity: (1) The content domain must be rooted in behavior which has a generally accepted meaning. The mere act of performing many tasks (e.g., driving an automobile) is generally accepted as meaning that the doer has the ability necessary to do them. (2) The content domain must be defined unambiguously; it ought to be possible for people who disagree on what the definition is to agree on whether a certain behavior is within or outside of the domain one of them has defined. (3) The content domain must be relevant to the purposes of the testing. That is, it should be a sample of some content domain external to the test--such as a job content domain--which can be identified in terms of testing practices such as selection, certification, or whatever. (4) Qualified judges must agree that the defined test content domain has in fact been adequately sampled; the procedures surrounding Lawshe's Content Validity Ratio (Lawshe, 1975) show one way to satisfy this requirement. (5) The response portion of the test content must be reliably observed and evaluated. To these five, I would now add the requirement that qualified judges agree that opportunities for contamination in the evaluation of responses be slight.

Fairness in Content Samples

The general acceptance of the idea of content domain samples as inherently fair is probably based on a line of reasoning somewhat like this: (1) The job content domain is independent of the characteristics of the people who hold the job. (2) A parallel test content domain and its sample will also be independent of such characteristics. (3) The standards for the evaluation of performance on the sample parallel those for performance on the job itself.

Bias in Job Content

The first two of these may be taken together; they do in fact pose some questions of fairness. For example, one consequence of affirmative action programs in many organizations seems to have been the creation (whether inadvertently or deliberately) of qualitatively different jobs for men and women or for minority and nonminority employees. Even where this happens, however, there is no necessary test unfairness attributable to it. It still seems likely that such jobs do in fact overlap in significant duties and responsibilities. The defined job content domain, as distinct from a larger job content universe,may therefore be the same in both groups of employees, and a common sample as a test is fair. If, on the other hand, the important and testable aspects of the two jobs differ greatly, then a test that is a sample from one is clearly unfair to the other--but the unfairness

lies not in the testing but in the creation of different jobs for men and women.

There is also a more subtle sense in which the assumption of independence of the job from personal characteristics can be questioned. (It must be recognized that I have no data here; ever mindful of the fact that these comments may someday be introduced in a court as an argument, I must emphasize that they are sheer speculation.) It has long been recognized that some jobs, notably management jobs, are partially defined by the styles of the people who hold them. Even on jobs below the management level, people differ in the ways they do a job and, in time, these differences may lead to differences in the jobs themselves.

I refer to this as "drift" in job definition. If stylistic differences are identified with racial differences, a cultural drift may occur. In what is supposed to be an increasingly integrated society, we seem paradoxically to find more and more voluntary social isolation of minorities. As members of one cultural group interact at work more with others like themselves and less with those in different cultural groups, the stage is set for drift in different directions leading to differences between cultural groups in their approaches to the job. A parallel drift, which has no connotation of race or sex, occurs for people who use different competencies in achieving the same ends.

If these differences are trivial in relation to the defined job

content domain, and if the domain as defined is important, then there is no more than a trivial issue of fairness. If the differences are indeed within the definition of the job domain, however, a phenomenon of "drift", if verified with data, could take on one of two diametrically different implications for fairness. First, if the defined job content domain is closer to the cultural experience of one group than to that of another, is it already unfair? I think not. This sort of difference is remarkably similar to some of the considerations behind the concept of bona fide occupational qualifications.

However, if the defined job content domain includes purely stylistic elements that are not necessarily relevant to the quality of performance, then I would suggest that the domain itself is erroneous and therefore unfair; any resulting test would be biased as the domain is biased. Job analysis techniques that emphasize style over the substance of duties and responsibilities may in fact pose a threat to fairness in testing if they lead to test content that requires a stylistic rigidity not inherent in the job.

Having said all of this, I must affirm the belief, unsubstantiated to be sure, that content domains, job or test, will rarely prove to be the culprit in unfair testing practices. In the first place, task-oriented, not worker-oriented, job analysis is the necessary foundation for content sampling in test construction (Prien, in press). In the second place, job analysis and domain definition will undoubtedly be done more carefully as more practitioners work with one eye on the

courts. The principle of job-content sampling is too intuitively appealing and too lacking in opportunity for statistical doubletalk, and the distinction between style and task is too dubious, for the general acceptance of inherent fairness to be lost solely because of a question about the independence of domain definitions from incumbent characteristics.

Bias in Performance Ratings

Rather, it is in the scoring of a content sample that the questions of subgroup differences must be faced. Consider a probationary period as a job content sample. Suppose that all comers were hired for a job and that they were given a series of probationary assignments faithfully representing later assignments. Certainly, we would call this a valid sample of the job. The problem is that performance on that valid sample must be evaluated.

Suppose that evaluation is obtained with the same psychometric method used in evaluating job performance after probation. If the probationary period is evaluated using supervisory ratings parallel to ratings used later, we find all of the familiar problems of ratings: the classic errors of halo, leniency, or central tendency; sources of systematic but unwanted variance such as length of acquaintance; the differences in bases of rating from one rater to another; unknown contaminations including ethnic or sex bias. Despite the injunction of the APA Standards to pay "particular attention" to these things, the

fact is that we do not really know how to evaluate them. Suppose, for example, the ratings in a criterion-related validity study are highly and positively related to length of service with the organization. Does that mean that the raters are biased in favor of the workers they've had around a long time and against the newcomers, or does it mean that experience really does count? Similar questions, of course, arise when ratings--or other criteria, for that matter--are significantly related to sex or ethnic classification. Is the relationship due to bias or to real differences in performance?

With ratings of probationary performance, all of these same questions haunt us, with the possible exception of the length of service issue. Without claiming a general solution, I can describe briefly some of our work on probationary performance evaluation of police officers by ratings obtained from training officers in the field. First, we are trying to separate description from evaluation; second, we are trying to spread the descriptive activity over a long period of observation. During this time, each observer records at least two <u>behavioral</u> performance observations each day: the best thing the probationer has done, and the worst thing, that day. Statements for a group of probationers are assembled and mixed, and the observers as a group then use Thurstone-type scaling to assign scale values to each observation. Third, we are training raters to be discriminating and precise in recording salient behaviors. Some training precedes the observation period; after the first two-week period,

each observer is shown the variances in the scaling of his statements so that he learns in his own written work the differences between glittering generalities and unambiguous descriptions. The rating for each probationer is the average of the scale values of the items written about him after this training is complete.

Will this elaborate and rather tiresome procedure eliminate bias? Probably not, but it does make a serious effort to minimize bias. Is making a good effort enough? Maybe not--but one thing is certain: if you don't make a very strong effort to minimize bias and to maximize objectivity in rating, subjective bias of one kind or more is sure to be a major if undetected source of variance in the ratings.

A measure is said to have construct validity to the degree to which the variance in a set of such measures can be systematically attributed to the construct being measured. How can we determine the construct validity of these probationary ratings?

The fact is that we can't do it very well. In time--in a period of time too long to be of any practical value--we could establish a kind of nomological network of information about the quality of performance on that job. If the probationary ratings fit logically in that network, such as it is, then we will have a basis for a judgment of construct validity in the measurement of the quality of performance in probationary assignments. It is important to notice that <u>that</u> judgment is independent of the earlier judgment of the representativeness of the sample of task making up the probationary assignments. A

probationary period lacking representativeness in the sampling of job tasks may nevertheless lead to valid ratings of performance. Conversely, a very carefully developed sample of the total job can be planned for probationary assignments, and it can be spoiled by invalid evaluations of performance on those assignments.

Bias in Scoring Tests

As a rule, of course, discussions of content samples do not focus on probationary periods. They are more likely to focus on work sample or job knowledge tests. It is important, therefore, to consider the measurement bias problems for these kinds of content samples.

A work sample test is likely to be scored in one of two ways: either the process of doing the work will be observed and evaluated, or the product will be evaluated (Shimberg, Esser, & Kruger, 1973). Consider first the evaluation of the product. It may be possible to evaluate it quite objectively; by weighing it, by subjecting it to systematic stress analysis, by determining the conductivity of solder connections, or by checking measurements. If such objective "scores" can be developed, and if their salience is accepted by a panel of qualified judges, then the evaluation of the work sample qualifies as one of those instances in which the measure can be accepted as an operational definition of skill in performance without further empirical validation.

If, however, the product must be judged without objective measure-

ment, or if there is no consensus about the appropriateness of the objective measures, problems of construct validity arise, independently of judgments of the adequacy (or validity) of content samples. Specifically, the problem is one of contamination from irrelevant sources of variance.

Policy-capturing research offers a useful approach to the problem. Each judge is studied (i.e., validated) independently. A substantial number of products of work samples is presented to the judge; a smaller set of variables is identified as the set of elemental characteristics of the product. Regression analyses can identify the characteristics weighted by each judge, and the pattern of their use, in developing a mathematical model of the judgments. The salient point is that possible contaminants may be included among the variables in the regression equations. As a simple illustration, if the judge knows the race and sex of the people behind the products, then race and sex--surely irrelevant to the evaluation of the product--can be entered into the equations. Evidence of bias, and therefore evidence against an interpretation of construct validity of the judgments, exists when such irrelevant variables have substantial beta weights or usefulness indices. Or, alternatively, a judgment policy or model can be based solely on relevant product characteristics and then used to predict actual judgments. Predicted judgments can be compared to actual judgments. Any consistent over- or underprediction for subgroups identified by the irrelevant variables is evidence of

contamination in the judgments.

If process is being judged (examples include drivers' examinations, check pilots, or evaluations of student teachers), similar policy-capturing research can be done. (We are currently modeling interviewer judgments using videotaped interviews.) Again, the inclusion of irrelevancies in the model, or the fact of consistent prediction errors using a relevant model, can serve as evidence of bias; conversely, the omission of the irrelevant characteristics from the model, or randomness of error in prediction can be used to support a case for the construct validity of the judgments.

Even for ordinary job knowledge tests, one may use the construct validity logic by testing hypotheses about irrelevant contaminating sources of variance. A multiple-choice examination given for occupational certification or licensing, for example, might be challenged because reading ability contributes to the variance in scores but does not have a place in the job content domain. This can be seen, of course, as a charge against content validity; i.e., the test content domain includes a task--reading--that is not included in the job content domain. Nevertheless, for the sake of the broader principle, let's recognize it as a charge of irrelevant variance in the scores. That is the fundamental nature of the question of fairness in the use of the test in any subgroup--or for any individual--where ability to read is likely to be impaired. (I'm not referring here to extreme impairment such as blindness; a better example is

the sort of impairment associated with poor educational foundations.) I see two approaches to the investigation of this source of variance. The more direct one is to correlate scores on the job knowledge test with scores on a standard test of reading comprehension; a low correlation supports the use of the job knowledge test. A quicker approach is to determine its readability level; if this is below some acceptable level, probably a level determined by consideration of basic literacy requirements or of job descriptions, the test may be accepted.

Whatever the approach, we must not lose sight of the basic problem. The basic problem is one of appraising possibly unfair use of a test. The fairness literature, emphasizing criterion-related validities, has,at least since Thorndike (1971),considered unfairness to result from irrelevant sources of variance in test scores. This is the broad problem in any determination of validity, including the validities of scores on tests justified as content samples. It must be considered whenever there is a reasonable suspicion that the numbers used to evaluate performance on a content sample are contaminated by sources of variance having nothing to do with the quality of performance. Properly, the variance in such scores would arise solely from such performance quality (with, of course, some permissible level of random error). So my fifth, and most important point is this: <u>Reasons for suspecting that contaminating sources of variance may be contributing to unfairness in the use of the scores, should be checked out--not ignored</u>.

I do not want to be misunderstood: I am not saying that every content sample should be evaluated by an elaborately scientific study of its construct validity; I am not contradicting earlier statements to the effect that content sampling can yield operational definitions of variables which are satisfactory solely for the method of their construction. Nor am I saying that empirical research should follow every irresponsible charge of unfairness. What I am saying is that psychologists cannot sit back on their content samples and assume that scores on content measures are fair merely because the content tasks fairly sample an appropriately defined domain.

A Mixture of Validities

None of this has, of course, talked about the "ability-performance relationship"; it has talked only of validity in the evaluative measurement of a sample of performance. Nevertheless, the topic intrigues me, and I want to comment on content sampling as a path of moderator variables for criterion-related validation.

Essentially, I'm building here on ideas presented at the Hawthorne symposium (Guion, 1975). There I urged the considerations of taxonomies of people based on polythetic rather than monothetic characteristics. To illustrate: In a doctoral dissertation at Bowling Green, Glenn Ball is developing a taxonomy based on five variables extracted from a self-description questionnaire. The questionnaire was, it should be added, completed as a part of the application for employment,

and he has a sample of some 8000 candidates. By the time he completes his dissertation, he will have a basis for classifying people in terms of profiles of measures of tolerance for annoying circumstances, spare time activities, and interpersonal attitudes. Subsequently, we shall use these profiles for assigning candidates to subgroups for studies of differential prediction.

This can be extended. People come to an employment officer with a background collection of skills. Within a clerical classification, for example, applicants may differ in terms of specifically job-related skills such as typing and shorthand, but they may also differ in terms of more general skills or knowledge: various kinds of arithmetic skill, skills in artistic production (such as bulletin-board layout), knowledge of business principles or of current events, and so on. I propose that research be undertaken to identify such broadly based content domains of culturally general knowledge and skill, to develop samples of these domains, and to develop taxonomies of people based on their characteristic profiles. Of course, such classifications may be predictive in their own right. They may, however, be related to many matters of current social concern, such as ethnicity, and since they are more fundamentally psychological in nature, they may prove to be the basis for effective moderator classifications.

A Parting Word

Obviously, I have no data on this last suggestion, and there is therefore no merit in pursuing it furtner at this time. The idea of a content sample as a moderator is, I trust, being offered and considered in moderation as well. The fact is the basis for my title. Whether we think of content samples as moderators, predictors, or criteria; whether we think of them as instruments for selection or for certification; whether we think of them as stop-gaps until the challenges to aptitude measures can be stilled or as the great hope for advance in mental measurement; whether we like them or not, content sampling is an important basis for test development. The care taken in defining content domains and in justifying them in relation to test purposes and in sampling from them should undergird many forms of testing beyond the work sample and the job knowledge test. But carefulness in test construction or content sampling should not be mistaken for validity.

I have said there is no such thing as content validity. Others have said it before me. I am sure the statement is true. Nevertheless, I am not naive enough to assume that people will stop talking about content validity, which does not exist, or about a content validation method of research, which therefore cannot exist. The term, content validity, will certainly stay in our vocabularies, probably to haunt us. Let us at least use the term in moderation, if not in moderating, and let us not attribute to it any magical power to eliminate problems of unfairness.

Reference Note

1. Jones, D. P. An examination of six fair employment models. Unpublished M.A. Thesis, Bowling Green State University, 1974.

References

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Division of Industrial-Organizational Psychology. Principles for the validation and use of personnel selection procedures. Dayton, Ohio: Division of Industrial-Organizational Psychology, American Psychological Association, 1975.

Ebel, R. L. Must all tests be valid? American Psychologist, 1961, 16, 640-647.

Ebel, R. L. Prediction? Validation? Construct validity? Personnel Psychology, 1977, 30, 55-63.

Guion, R. M. The Hawthorne type--among others. In E. L. Cass & F. G. Zimmer (Eds.), Man and work in society. New York: Van Nostrand & Reinhold, 1975.

Guion, R. M. Recruiting, selection, job placement. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Guion, R. M. Content validity--the source of my discontent. Applied Psychological Measurement, 1977, 1, 1-10.

Hull, C. L. Aptitude testing. Yonkers-on-Hudson: World Book Co., 1928.

Lawshe, C. H. A quantitative approach to content validity. Personnel Psychology, 1975, 28, 563-575.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Munsterberg, H. Psychology: General and applied. New York: Appleton, 1915.

Office of Federal Contract Compliance. Validation of tests by contractors and subcontractors subject to the provisions of Executive Order 11246. Federal Register, September 24, 1968, 33 (No. 186), 14392-14394.

Prien, E. Job analysis in content validity. Personnel Psychology, in press.

Shimberg, B., Esser, B. F., & Fruger, D. H. Occupational licensing: Practices and policies. Washington, D.C.: Public Affairs Press, 1973.

Spearman, C. The abilities of man: Their nature and measurement. New York: Macmillan, 1927.

Tenopyr, M. L. Content-construct confusion. Personnel Psychology, 1977, 30, 47-54.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Moderator Research and the Law of Small Numbers

Frank L. Schmidt

U.S. Civil Service Commission and George Washington University

John E. Hunter

Michigan State University

Abstract

The thesis of this paper is that many proposed moderators in personnel psychology are probably illusory, having been created solely by belief in the law of small numbers. Evidence is presented that race as a moderator of test validity is one such illusory moderator. In addition, a model for validity generalization is described which, in addition to eliminating the need for criterion-related validity studies under certain circumstances, strongly calls into question the idea that situations moderate test validity, i.e., the traditional doctrine of situational specificity of test validities. Calculations are presented which show that adequate statistical power in moderator research requires much larger sample sizes than have typically been employed. This requirement is illustrated empirically using validity data for the Army Classification Battery for 35 jobs and 21,000 individuals. These analyses show that (1) even when a moderator is generally assumed to be large, large samples are required to gauge its effect reliably and (2) large sample research may show that moderators that appear plausible and important a priori are nonexistent or trivial in magnitude. The practice of pooling across numerous small sample studies to obtain statistical power equivalent to that of large sample studies is recommended. In light of the evidence that many proposed moderators may not exist, the authors hypothesize that the true structure of underlying relationships in personnel psychology is considerably simpler than personnel psychologists have generally imagined it to be.

Moderator Research and the Law of Small Numbers

Frank L. Schmidt

U.S. Civil Service Commission and George Washington University

John E. Hunter

Michigan State University

Let us start with a clear statement of our biases. In our opinion, it is highly likely that belief in the law of small numbers (Tversky & Kahneman, 1971) has in the past led not only to methodologically poor moderator research but also to the acceptance of illusory moderators. We think it is likely that rigorous research will some day show that the number of moderators that actually exist in populations, as opposed to samples, is quite small. The underlying pattern of relationships among population parameters in personnel psychology may be much simpler than we have hypothesized it to be.

The law of small numbers holds that small random samples can be considered to be about as representative of their corresponding populations as large random samples are. The believer in the law of small numbers habitually overestimates the amount of information about the population contained in small samples and the power of statistical analysis to extract that information. As a result, he frequently employs small samples in his research, samples that can under no circumstances answer the questions put to them. Then when such research produces results inconsistent with predictions, as it must for sta-

tistical reasons alone, additional "explanatory variables" -- often moderators -- are hypothesized, which might account for what in actuality was produced by statistical artifacts.

Belief in the Law of Small Numbers in Action

Let us now examine examples of this tendency in action. After passage of the 1964 Civil Rights Act, the hypothsis was advanced that employment tests have different validity coefficients for whites than blacks. The alternative hypothesis, of course, was that no such differences existed. This alternative was not only more parsimonious, but as Humphreys (1973) has pointed out, has a stronger conceptual and theoretical foundation. After all, blacks and whites live in the same society, watch the same television shows, attend similar schools, etc. On the other hand, no systematic rationale for differential validity was ever presented. No one in personnel psychology ever bothered to present a theoretical mechanism by which black-white cultural differences, if they exist to any extent (beyond social class differences), would specifically affect validity coefficients. The differential validity hypothesis was nevertheless immediately popular with many, if not most, industrial psychologists.

The next step, of course, was to test the hypothesis in small sample studies. Methodologically, these studies fell into two categories: single group validity studies and differential validity studies. In the former both black and white validity coefficients are tested for statistical significance. In the latter, the signifi-

cance of the difference between the coefficients is tested. It is now well known that the single group validity procedure is erroneous, even apart from the sample size question[1] (Humphreys, 1973; APA, 1974, p. 43).

Since minority sample sizes were usually smaller than those for the majority, small sample single group validity studies produced a high frequency of white-significant, black-nonsignificant findings. When we (Schmidt et al., 1973) showed that the findings of all single group validity studies could be accounted for by chance alone under the assumption of equal black-white validities, many psychologists personally told me they did not believe the results. They had seen so many empirical validity coefficients that were not identical in size or significance for blacks and whites that they could not entertain the idea that error alone might produce such variations. Despite the absence of any theoretical rationale for race as a moderator, the idea of race as a moderator had more appeal than the more parsimonious and more theory-based explanation in terms of statistical artifacts. As a result, they simply could not accept our findings. Some even continued to do small sample single group validity studies.

The differential validity approach, on the other hand, is statistically sound, but unless true validity differences are very large,

[1]For a discussion of other statistical and methodological errors in this area, see Hunter and Schmidt, in press.

the chances of detecting them are not good unless very large samples are used. For example, if true validities are .30 and .50 for blacks and whites, respectively, and criterion reliability is .70, then even in the absence of any range restriction at all, one must have 528 in each group to have a .90 probability of detecting this difference (two-tailed test, Alpha = .05). The sample sizes typically used in any one study are thus inadequate to provide a meaningful test of the hypothesis.

One can, however, attain high statistical power by combining across studies. We have done this in recent research which included the results from some 33 studies producing over 800 pairs of validity coefficients. The frequency of differential validity by race was found to be at chance level (Hunter, Schmidt, & Hunter, Note 1). And yet we feel confident that many will not accept the obvious conclusion. The less parsimonious and less theoretically justified hypothesis will continue to be entertained and many researchers will continue to believe that the highly untrustworthy coefficients they compute in their small samples can tell them much more about the truth or falsity of the validity-difference hypothesis than the weight of accumulated evidence.

Industrial psychologists have long been ardent believers in another, vaguer and less well defined moderator of validity: the undifferentiated situation. The belief that test validity is generally highly situation-specific has long been one of the central orthodoxies

in industrial psychology. Considerable variability from study to study is in fact observed in raw validation results even when jobs and tests appear to be similar or essentially identical (Ghiselli, 1966). The orthodox explanation for this phenomenon is that the factor structure of job performance is different from job to job and that the human observer or job analyst is simply too poor an information receiver and processor to detect these subtle but important differences. That is, there are mysterious, unknown and maybe unknowable moderator variables operating, causing validity to be high in one setting and low or zero in another. Therefore, it is concluded, empirical validation is required in each situation, and validity generalization is impossible (Ghiselli, 1966, p. 28; Guion, 1965, p. 126; Albright, Glennon, & Smith, 1963, p. 18). This harsh "fact" is widely lamented, and it is said that our inability to solve the problem of validity generalization is perhaps the most serious shortcoming in selection psychology today (Guion, 1976; Division of Industrial-Organizational Psychology, 1975). The inability to generalize validities precludes development of general principles of selection that could take our field beyond a mere technology to the status of a science (Guion, 1976).

But there is evidence suggesting that much of the variance in the outcomes of validity studies within job-test combinations may be due to statistical artifacts. Schmidt, Hunter, and Urry (1976) have shown that under typical and realistic validation conditions, a valid

test will show a statistically significant validity in only about fifty percent of studies. As one specific example, they showed that if true validity for a given test is constant at .45 in a series of jobs, if criterion reliability is .70, if the prior selection ratio on the test is .60, and if sample size is 68 [the median over 406 published validity studies (Lent, Aurbach, & Levin, 1971)], then the test will be reported to be valid 54 percent of the time and invalid 46 percent of the time (two-tailed test, $p < .05$). These are the kinds of results that are in fact observed in the literature (Ghiselli, 1966). When sample size is adequate to provide appropriate levels of statistical power, the observed results are quite different. In a well executed large sample series of studies, it was found that when Army occupations were classified rationally into job families, tests showed very similar validities and regression weights for all jobs within a given family (Brogden, Note 2). Further, new jobs assigned rationally to job families also fit this pattern. Brogden has concluded that when methodological artifacts are controlled and large samples are used, obtained validities are in fact stable and similar across time and situations for similar jobs.

If the variance in validity coefficients across situations for job-test combinations is due to statistical artifacts, then obviously the doctrine of situational specificity is false and validities are generalizable. We have developed a method for testing this hypothesis (Schmidt & Hunter, in press).

This method can be explained conceptually as follows. One starts with a fairly large number of validity coefficients for a given test-job combination. These are then converted to Fisher's Z and the variance of this distribution is computed. From this variance, one then subtracts variance due to various sources of error. These sources include:

1. Small sample sizes (i.e., variance induced by the less than infinite sample sizes in each of the studies).
2. Differences between studies in criterion reliability.
3. Differences between studies in range restriction.
4. Computational and typographical errors.
5. Differences between studies in amount and kind of criterion contamination and deficiency (Brogden & Taylor, 1950).

If, after subtracting variance due to these sources, the variance of the distribution of the validity coefficients is essentially zero, the hypothesis is confirmed.

Even if the remaining variance is not zero, there may still be important implications. After correcting the mean of this distribution for attenuation due to criterion unreliability and for range restriction (based on average values of both), it may become apparent that a very large percentage, say 95 percent, of all values in the distribution lie above the minimum useful level of validity. In such a case, one could conclude with 95% confidence that true validity was at or above this minimum level in a new situation involving this test-

type and job without carrying out a validation study of any kind. Only a job analysis would be necessary--to insure that the job at hand was indeed a member of the class in question.

Thus validity generalization can be justified even where it cannot be shown that all of the observed variance is error variance. In a sense, this outcome does not refute the hypothesis of validity specificity: validities can still differ somewhat from situation to situation. But the most (and only) serious consequence of validity specificity--the inability to generalize--is circumvented. In addition, as we shall see in a minute, little room may be left in which situational moderators can operate.

How does one proceed in correcting observed variance for error variance due to the sources listed above? First, consider variance due to sample size. In this case, one need only know the average sample size across published studies. A recent review of 406 studies (Lent, Aurbach, & Levin, 1971) places this figure at 68. Variance due to sample size can therefore be estimated as $1/N-3$ or 65^{-1}. This estimate is conservative since published studies tend to average higher sample sizes than unpublished studies (Guion, 1965, p. 126). In the case of variance due to differences between studies in criterion reliability, one assumes a reasonable distribution of reliabilities across studies and then determines the amount of variance this distribution would contribute to the observed distribution of validities. Table 1 shows such an assumed distribution of criterion reliabilities.

Table 1

Example of Assumed Distributions of Criterion Reliabilities Across Studies

Reliability	Relative Frequency
.90	3
.85	4
.80	6
.75	8
.70	10
.65	12
.60	14
.55	12
.50	10
.45	8
.40	6
.35	4
.30	3

E (Reliability) = .60

The same procedure is followed in the case of differences between studies in range restriction; an example of such an assumed distribution of range restriction effects can be seen in Table 2. In the case of both criterion reliability and range restriction, the information necessary to determine actual values in individual studies is not presented in the vast majority of research reports (Jones, 1950). Thus one must rely on reasonable assumed distributions of these effects; for reasons given later, these distributions should usually be conservative. The procedures by which one computes estimates of variance due to criterion reliability and range restriction effects are given in Appendix A. After computation, all three of these variances are subtracted from the observed variance, providing the final estimate of true situational variance, i.e., variance due to true differences between tests and jobs. Note that no correction has been made for differences between studies in amount and kind of criterion contamination or deficiency or for computational and typographical errors. Although computational and typographical errors are probably more frequent than usually assumed (Wolins, 1962), it is difficult to estimate their frequency or magnitude and thus difficult to correct for them. In the case of criterion deficiency or contamination, corrections would be even more difficult. In addition, not correcting for these sources of error insures a "conservative" procedure in that the corrected variance tends to overestimate situational specificity.

A computer program which makes the corrections described above

Table 2

Example of Assumed Distribution of Range Restriction Effects Across Studies

Prior Selection Ratio	SD of Test	Relative Frequency
1.00	10.00	5
.70	7.01	11
.60	6.49	16
.50	6.03	18
.40	5.59	18
.30	5.15	16
.20	4.68	11
.10	4.11	5

E (SD) = 6.0

was written and applied to four validity distributions presented by Ghiselli (1966, p. 29). These distributions contain both published and unpublished validity coefficients. Application of the D'Agostino and Cureton (1972) test, the most powerful such test available, showed that the distributions in Fz form did not depart significantly from normality, thus allowing use of the normal model with these data. Criterion reliability and range restriction inputs were those shown in Tables 1 and 2, respectively.

The results of central interest are shown in Table 3. For general clerks and mechanical repairmen, the corrected distributions (priors) are such that one can be virtually certain that the tests in question are valid in a new setting without carrying out a criterion-related validity study. For the mechanical repairmen, we are 97.5 percent confident that true validity is .70 or higher, and for the clerks we can have this same degree of confidence that true validity is at least .40. In the case of the bench workers, the 97.5% credibility interval includes zero. The 95% confidence interval (not shown in Table 4) goes down to r = .03. Thus this distribution does not allow conclusions about validity in the absence of a study. The same is true in the case of the machine tenders. This prior is exactly opposite to that for the mechanical repairmen. In this distribution, only 7 percent of the coefficients lie above .30. This result suggests the hypothesis that the true validity of spatial relations tests for machine tenders is zero and that the remaining variance is due to

Table 3

Results of Pilot Study

Job	Test Type	Number of Validity Coefficients	Prior Distribution Mean r*	Mean Fz*	SD_{Fz}	97.5% Credibility Value for Validity
Mechanical Repairman	Mechanical Principles[a]	114	.78	1.03	.08	.70
Bench Workers	Finger Dexterity[b]	191	.39	.41	.23	-.04
General Clerks	Intelligence[b]	72	.67	.81	.20	.40
Machine Tenders	Spatial Relations[b]	99	.05	.05	.18	-.30

*Corrected for range restriction and attenuation due to criterion unreliability.

[a]Training criteria

[b]Proficiency criteria

Table 4

Expected SD's of Fisher's Z Validity Distributions Given No True Differences Between Jobs and Tests

Job--test type combination	(1) Range Restriction	(2) Criterion Unreliability	(3) (1) and (2)	(4) (1) and (2) and N=30	(5) (1) and (2) and N=50	(6) (1) and (2) and N=68	(7) Observed SD
Mechanical Repairman, Mechanical Principles[a]	.081	.080	.142	.239	.203	.189	.205
Bench Workers, Finger Dexterity[b]	.037	.041	.056	.200	.157	.136	.262
General Clerks, Intelligence[b]	.068	.088	.111	.222	.184	.167	.263
Machine Tenders, Spatial Relations[b]	.005	.005	.007	.193	.146	.124	.219

[a]Training Criteria

[b]Proficiency Criteria

sources of artifactual variance not controlled for.

Results like those for the mechanical repairmen and general clerks show that validity generalization is possible. But since the residual variance is not zero, these results do not completely preclude some degree of situational specificity. Conceptually anyway, some of the remaining variance may be due to unknown situational moderator variables. Let us examine the likelihood that this is in fact true. Table 4 shows the SD's of validity distributions that would be observed if there were in fact no true validity differences and all observed variance was due to various artifactual sources and combinations of sources. These figures can be compared with the observed SD's (column 7). In the case of the mechanical repairmen, for example, given only criterion reliability and range restriction differences between studies and an average sample size of 68, the expected SD is .189 (column 5). This compares with an observed SD of .205. Thus these three artifactual sources of variance account for 85 percent of the observed variance. If we assume the more realistic average sample size of 50 (Guion, 1965), the expected standard deviation is .203, almost identical to the observed standard deviation of .205. Thus 98 percent of observed variance is accounted for. In the case of general clerks, the assumption of a mean sample size of 68 leads to a predicted SD of .167, as compared to the observed value of .263. Forty percent of the observed variance is accounted for. If the mean sample size is taken as 50, our three statistical artifacts account for 48 percent

of observed variance.

In the case of both distributions, but especially in the case of mechanical repairmen, the amount of remaining variance is quite small. In addition, sources of error variance not corrected for in our model almost certainly account for some of this residual variance. As pointed out earlier, no correction has been made for differences between studies in amount and kind of criterion contamination and deficiency. Nor has any correction been made for typographical and computational errors.

In moderator research, unlike test validation, an additional artifactual source of variance becomes relevant. Tests of a given general type, for example mechanical principles tests, differ from each other in validity at least somewhat, for at least two reasons. First, they vary somewhat in reliability, which, even given identical factor structures, produces variation in validities. Secondly, factor structures differ at least slightly from test to test. It is not legitimate to correct for this source of variation when validity generalization is one's focus. Validity generalization, when it is justified, is made to the group of tests available to the applied psychologist for possible use. The variation we have just discussed is inherent in that population of tests and must be included in the distribution on which validity generalization is based. Moderator research, however, is theoretical research. As such, it is concerned with relations among underlying constructs, independent of measurement

problems. Thus from the viewpoint of moderator research, variance due to differences between tests in reliability and content is artifactual and should be partialed out.

How much variance would remain within which a moderator could operate if we could partial out not only the three artifactual sources we have partialed out but also the four we have not? Perhaps none. At best, very little. This means that the population moderator effect, if one exists, will in most cases be quite small. For example, the population correlation may be .55 in one group or condition and .60 in the other. Such a situation raises problems of statistical power that are well nigh insoluble in individual studies. Effects this small are virtually impossible to detect in any given study. This problem is discussed below.

Would the presence of such a moderator--whether we could detect it or not--preclude validity generalization? That is, would it lead to decision errors when generalizing validities? Generally, it would not. Suppose, again, that in one group or condition, true validity were .55 and in another group or condition it were .60. Suppose further that the group with the true validity of .55 made up 50 percent of the population. The corrected prior would then have an expected mean of .575, which is .65 in Fisher's Z form. Now suppose, further, that the corrected SD were .08, as in the case of the mechanical repairmen. That is, this SD of .08 reflects the effects of the four artifactual sources of variance we have not partialed out plus the between-group

variance produced by the moderator. The lower 95 percent credibility interval would be about .48 for the group as a whole. For the group with true validity of .60, it would be .51; for the other group, the lower bound would be .45.[2] If the SD of the prior were .20 instead of .08, the 95 percent credibility interval points would be .35 for one group, .30 for the other, and .31 for the total group. Thus even with larger SD's the presence of such moderators is not apt to lead to decision errors when generalizing validities.

A more important point is the fact that such moderators, if they did exist, would have very limited practical utility. Suppose that, despite the odds against it, one succeeded in demonstrating the validity coefficient were .05 points higher in one group or condition than another. In our first example, where the SD of the corrected prior is .08, this is actually a fairly substantial moderator effect: .07 Fisher's Z units, or 7/8 of a SD. And yet what is the practical usefulness of this knowledge? We really can't see any.

The gloomy scenario we have sketched here for research on situational moderators may reflect our own hunches and hypotheses as much as it reflects reality. After all, our validity generalization model

---

[2] In these examples, variance due to the moderator is .000625. Thus the within moderator group variance in the first example is $.08^2 - .000625 = .00575$. SD = $\sqrt{.00575} = .0779$. In the second example, the within moderator group variance is $.20^2 - .000625 = .039375$. SD = $\sqrt{.039375} = .19874$. These within group SD's are used in computing lower bounds of the credibility interval.

has thus far been applied to only four test-job combinations. The examination of other validity distributions in the future may reveal some in which the corrected variance remains so large that non-trivial moderator effects of some sort become genuine high-probability hypotheses. And some of these moderator effects, if they exist, may be large enough to have practical implications. These are possibilities, but the fact that impresses us most is that when even conservative allowances are made for sources of artifactual variance, the variance of validities is suddenly seen to be much smaller than most psychologists believe. Apparently, personnel psychologists have a strong tendency, maybe a psychological need, to interpret error variance as true variance. If we can strip this tendency away, we may find that reality is simply and elegantly structured.

Statistical Power in Moderator Research

If one's total group of subjects can be divided into two or more subgroups based on some classification variable, if the correlations between variables of interest are significantly different in different groups, and if these differences are not due to sample artifacts such as differential range restriction or differential reliability of one or both variables, one has evidence for the operation of a moderator. Note that what is required is a significant difference between the coefficients; it is not sufficient that one coefficient be statistically significant and the other not (Humphreys, 1973; American Psychological Association, 1974).

Let us examine the question of sample sizes required in moderator research to provide adequate statistical power in a given study to detect these differences. Total sample size requirements are minimized when the two groups contain equal numbers. If moderator effects are in fact trivial in magnitude, the sample size problem is indeed formidable. Example 1 in Table 5 shows that when one population correlation is .55 and the other is .60, both variables have reliability .70, and there is no range restriction whatsoever, then even if one can specify in advance the direction of the difference and is willing to settle for power of .80, he needs almost 8,000 subjects in each group. The situation brightens somewhat as the size of the moderator effect increases. If the true correlations are .30 and .50, as shown in Example 2 of Table 5, then a one-tailed test of power .80 requires "only" 539 subjects in each group. The total required sample size is thus over 1,000. Example 3 shows results for an even larger moderator effect. The true correlation is .20 in one group and .60 in the other. Here the sample size required for the same statistical test is 134 in each group, or a total sample of 268. Situational moderator effects this large are, in our judgment, apt to be exceedingly rare if they exist at all. Nevertheless, sample sizes even this large are typically not found in the moderator research literature.

Actually, there is, in theory at least, a method available to circumvent the need for high statistical power in individual studies and we will examine this in the final section of the paper. But first we

Table 5

Examples of Sample Sizes Required in Each Group for Two Levels of Statistical Power in Detecting Differences Between Correlations in the Absence of Range Restriction

(Reliability=.70 for Both Variables; Alpha=.05)

I. Example 1: $r_{(true\ 1)} = .55$, $r_{(true\ 2)} = .60$

	Power	
	.90	.80
2-Tailed	13,125	9,873
1-Tailed	10,661	7,753

II. Example 2: $r_{(true\ 1)} = .30$, $r_{(true\ 2)} = .50$

	Power	
	.90	.80
2-Tailed	912	686
1-Tailed	741	539

III. Example 3: $r_{(true\ 1)} = .20$, $r_{(true\ 2)} = .60$

	Power	
	.90	.80
2-Tailed	224	170
1-Tailed	183	134

would like to present some empirical data indicating the need for large sample sizes if meaningful conclusions about moderators are to be drawn from individual studies. These data are from research on the Army Classification Battery (Helm et al., Note 3). Table A-1 in the Appendix shows validity coefficients for the 10 subtests of this battery for training success measures in 35 army jobs. Sample sizes in these jobs ranged from 100 to over 500, with an average of 300. Table A-2 shows similar validity estimates computed on a different and independent set of samples. In this table, the average sample size across jobs is again 300. Each of the tables is based on 10,500 individuals.

An examination of these tables shows that the 35 jobs vary widely. There are, for example, welders, cooks, clerks, and administrators. Thus we would expect jobs to moderate test validity. That is, we would expect a given test to show reliable differences in validity for different jobs. (If such widely varying jobs do not moderate validity, it is hard to imagine situational variables that would.) As would be expected and desired in a classification battery, the ten tests in the battery also vary widely. The tests are described in Table A-3. Thus we would expect tests to moderate validity with respect to jobs. That is we would expect different tests to show different patterns of validity across the 35 jobs.

Table 6 shows cross-sample correlations of validity profiles for tests across jobs. These are correlations between data in Tables A-1 and A-2. Values on the diagonal indicate, for each test, the stability

Table 6

Correlations Across Sample Sets of Validity Profiles of Tests

	V	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
V	.71	.75	.48	.38	.67	.24	.37	.06	.33	.07
AR		.82	.66	.54	.62	.14	.49	.19	.54	.25
PA			.68	.68	.36	.06	.66	.46	.59	.34
MA				.76	.20	-.06	.73	.65	.63	.33
ACS					.71	.19	.26	.02	.11	-.15
ARC						.19	.19	.22	.05	.09
SM							.79	.72	.54	.18
AI								.85	.38	.03
EI									.86	.74
RI										.84

Note. Off diagonal correlations are averages (using Fisher's z) of both relevant correlations. For example, the correlation profiles for tests AR and V is the average of the correlation V-Sample Set A vs. AR-Sample Set B and V-Sample Set B vs. AR-Sample Set A.

of its validity profile across jobs. The value for Army Radio code is .19, but the others range from .68 to .86. Thus with an average sample size of 300 per job, it is possible to establish reasonably reliable validity profiles for tests. That is, it is possible to get reasonably reliable estimates of the extent to which jobs moderate test validity. The important point, however, is that these results were produced using sample sizes so large as to be virtually unattainable for the typical researcher. With the kinds of sample sizes that are typical, one would find it difficult in individual studies to accurately determine differences between jobs in test validity--even though the moderating effect of jobs would be expected to be large relative to that of situational moderators.

Our second hypothesis was that tests would moderate validity with respect to jobs. Table 7 shows cross-sample correlations between test validity profiles corrected for unreliability. (Reliabilities used were, of course, those on the diagonal of Table 6.) Surprisingly, the evidence strongly calls this hypothesis into question for some test pairs. For example, consider the Vocabulary and Arithmetic Reasoning tests. Most psychologists would probably consider it a very reasonable and plausible hypothesis that these two tests would show a different profile of validities across a set of job. They would probably feel that for each job, the specific duties, content, etc., would determine the relative validity of these two tests. But the data indicate otherwise. The estimated correlation between

Table 7

Correlations from Table 6 Corrected for Attenuation and Mean Validities Across Jobs

	V	AR	PA	MA	ACS	ARC	SM	AI	EI	RI	Mean Validities Sample A	Mean Validities Sample B
V	1.00	.98	.69	.52	.94	.65	.49	.08	.42	.09	.51	.52
AR		1.00	.88	.68	.81	.36	.61	.23	.64	.30	.56	.57
PA			1.00	.95	.52	.17	.90	.61	.77	.45	.47	.49
MA				1.00	.27	-.16	.94	.81	.78	.41	.51	.50
ACS					1.00	.52	.35	.03	.14	-.19	.39	.42
ARC						1.00	.49	.55	.12	.23	.34	.35
SM							1.00	.87	.65	.22	.48	.48
AI								1.00	.44	.04	.41	.38
EI									1.00	.87	.45	.44
RI										1.00	.32	.32

true profiles is about as close to 1.00 as possible. The evidence clearly indicates that profile shape is identical.

Table 7 shows many slightly lower correlations--correlations in the .80's and .90's. These correlations indicate pairs of tests in which validity profiles are highly similar but apparently not identical. But since they are so similar, i.e., since the moderator effect is so small, these differences would be virtually impossible to detect in empirical research. Thus, a moderator effect which a priori appeared important is probably often nonexistent or too small to be detected.

High correlations in Table 7 indicate the absence of the large "test by job interactions" which define practically significant and theoretically interesting moderator effects. They do not rule out test main effects or level differences between test profiles of highly similar or identical shape. Main effects are indexed by mean validity of tests across jobs, and these are also shown in Table 7. The Vocabulary and Arithmetic Reasoning have identical profile shapes, but the Arithmetic Reasoning test is generally about 5 correlational points higher in validity. These level differences are at present not interpretable. Validity profile level depends on test reliability. The Arithmetic Reasoning test may, for example, simply be more reliable than the Vocabulary test. Thus far, we have not been able to obtain reliability figures on these tests, although we may be able to in the future.

In conclusion, the Army data illustrate two points:

1. Even where a moderator effect is generally assumed to be large--like the effect of jobs on the validity of a test--large sample sizes are required to reliably and validly gauge the effect.
2. When the proper large sample research is done, moderators that appear plausible and important a priori may be shown to be nonexistent or trivial in magnitude.

Future Prospects for Moderator Research

Based on our remarks up to this point, one might assume that we see future prospects for sound moderator research as dark indeed. Although we do feel that these prospects are less than bright, we do not think the situation is hopeless. It is true that adequate statistical power in any given study requires sample sizes that are rarely attainable. But it is also true that "studies of studies," that is, studies combining the results of numerous studies, can have high levels of statistical power--even though component individual studies contain only small or modest N's (Hunter, Schmidt, & Hunter, Note 1). Although this procedure proved workable in the case of race as a moderator, it may be difficult for other proposed moderators to stimulate such large numbers of research studies. Past research patterns indicate a tendency for each researcher to focus on the moderator that particularly interests him. Except in the case of jobs as moderators, replications have been rare. But it is neverthe-

less possible that one or more proposed moderators will in the future capture the imagination of researchers as race did and thus lead to the generation of the reams of necessary data.

Moderator research is more apt to be successful if the moderators researched are not only nonzero but also nontrivial in magnitude. Both these conditions are more apt to be fulfilled when moderators are hypothesized on the basis of careful a priori theorizing than when the research basis is shotgun empiricism or even unarticulated hunch. Thus our feeling is that, if any future moderator research succeeds, it will probably be based on a careful a priori theoretical foundation which justifies postulating a relatively large moderator effect. In addition, the theoretical work will be followed up by either a fairly large number of small sample studies or a smaller number of large sample studies.

Our hunch is that many proposed moderator variables simply do not exist. We strongly suspect that the underlying structure of reality in personnel psychology--that is, the pattern of relationships among population parameters--is considerably simpler than we have imagined it to be. By reading error variance like tea leaves, we have conjured up visions of complexity. Many of these visions are illusions.

Reference Notes

1. Hunter, J. E., Schmidt, F. L., & Hunter, R. Differential validity of employment tests by race: A disconfirmation. Manuscript in preparation.

2. Brogden, H. E. Personal communication, 1970.

3. Helm, W. E., Gibson, W. A., & Brogden, H. E. An empirical test of shrinkage problems in personnel classification research. Personnel Research Board Technical Research Note 84, October, 1957.

References

Albright, L. E., Glennon, J. R., & Smith, W. J. The uses of psychological tests in industry. Cleveland: Howard Allen, 1963.

American Psychological Association, American Educational Research Association, and National Council of Measurement in Education. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Brogden, H. E., & Taylor, E. K. A theory and classification of criterion bias. Educational and Psychological Measurement, 1950, 10, 159-186.

D'Agostino, R. B., & Cureton, E. E. Test of normality against skewed alternatives. Psychological Bulletin, 1972, 78, 262-265.

Division of Industrial-Organizational Psychology. Principles for the validation and use of personnel selection procedures. Dayton, Ohio: Division of Industrial-Organizational Psychology, American Psychological Association, 1975.

Ghiselli, E. E. The validity of occupational aptitude tests. New York: Wiley, 1966.

Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

Guion, R. M. Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), Handbook of industrial-organizational psychology. Chicago: Rand McNally, 1976.

Humphreys, L. G. Statistical definitions of test validity for minority groups. Journal of Applied Psychology, 1973, 58, 1-4.

Hunter, J. E., & Schmidt, F. L. Differential and single group validity of employment tests by race: A critical analysis of three recent studies. Journal of Applied Psychology, in press.

Jones, M. H. The adequacy of employee selection reports. Journal of Applied Psychology, 1950, 34, 219-224.

Lent, R. H., Aurbach, H. A., & Levin, L. S. Predictors, criteria, and significant results. Personnel Psychology, 1971, 24, 519-533.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 53, 5-9.

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, in press.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validity studies. Journal of Applied Psychology, 1976, 61, 473-485.

Thorndike, R. L. Personnel selection. New York: Wiley, 1949.

Tversky, A., & Kahneman, D. Belief in the law of small numbers. Psychological Bulletin, 1971, 76, 105-110.

Wolins, L. Responsibility for raw data. American Psychologist, 1962, 17, 657-658.

Appendix A

1. Computing variance due to differences between studies in criterion reliability.

   1. Compute mean of the validity distribution in Fisher's Z ($\bar{F}$) form and convert to r.
   2. Correct this raw r for criterion unreliability and range restriction using average values across studies for both. (In the pilot study, average assumed criterion reliability was .60 and average range restriction was to a SD of 6.0 from an unrestricted SD of 10.0; see Tables 1 and 2 in text.) This provides an estimate of the mean true validity, r_{oo}.
   3. For each value of assumed criterion reliability, r_{tt_i}, compute $r_{oo}\sqrt{r_{tt_i}}$ and convert this attenuated r to Fz. Compute $\Sigma Fz_i \cdot n_i$ and $\Sigma Fz^2{}_i \cdot n_i$, where n_i = the relative frequencies of the criterion reliabilities.
   4. Variance due to reliability differences in Fz distribution of validities is then:

$$\sigma^2_{r_{tt}} = \frac{\Sigma Fz^2{}_i \cdot n_i}{\Sigma n_i} - \left[\frac{\Sigma Fz_i \cdot n_i}{\Sigma n_i}\right]^2$$

II. Computing variance due to range restriction differences between studies:

1. Compute mean of the validity distribution in Fz form and convert to r. Correct this raw r for mean range restriction but not for attenuation.
2. For each value of the restricted standard deviation, use the following formula to compute the expected restricted r:

$$r_i = \frac{u_i R}{\sqrt{u^2_i R - R^2 + 1}},$$

where:

r_i = the restricted validity

R = the unrestricted validity

u_i = sd_i/SD

SD = the standard deviation of the test in the unrestricted group

sd_i = the standard deviation of the test in the restricted group

This formula is obtained by solving Thorndike's (1949, p. 173) Case II formula for r_i. (Thorndike's Case II is the model throughout these analyses; use of Case III would generally produce very similar results.)

3. Convert r_i to Fz and compute $\Sigma Fz_i \cdot n_i$ and $\Sigma Fz^2_i \cdot n_i$.
4. Variance due to range restriction differences between studies is then:

$$\sigma^2_{rr} = \frac{\Sigma Fz^2_i \cdot n_i}{\Sigma n_i} - \left[\frac{\Sigma Fz_i \cdot n_i}{\Sigma n_i}\right]^2$$

Table A-1

Validity Coefficients* of ACB Tests for 35 Army School Courses

Group A. Samples

Current MOS	Course Title	N	ACB Tests									
			RV	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
271-2	Fixed Sta Radio Rep	310	58	69	58	61	38	52	57	42	67	62
281	Microwave-Mult Channel Rep	216	36	52	41	48	23	16	44	40	55	54
282	Radar Rep	242	38	46	40	55	22	24	32	38	50	48
296	Field Radio Rep	280	54	62	60	57	39	42	59	50	63	48
320	Field Wireman	330	45	49	48	43	31	32	52	43	58	40
403	FC Instrument Rep	214	52	49	58	59	34	31	54	50	52	37
411	Ammunition Supply	767	54	55	45	42	44	28	50	42	38	20
421	Arty Mech-Light Weapons	196	64	64	57	58	54	29	63	60	49	13
422	Small Arms Weapons Mech	418	53	60	60	61	48	34	69	64	44	26
424	Turret Arty Rep	183	66	70	60	62	52	41	66	61	44	28
442	Welder	236	55	50	59	59	37	33	56	56	38	19
443	Machinist	296	38	59	57	58	33	30	66	46	46	34
452	Dental Lab	100	38	44	52	60	24	29	53	40	40	28
631	Automotive Mech	154	31	48	45	57	28	17	51	64	45	28

Table A-1 (continued)

Current MOS	Course Title	N	ACB Tests									
			RV	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
632	Track Veh Rep	448	62	64	56	67	47	34	69	69	49	27
632	Armor Track Veh Maint	248	48	56	54	65	36	36	58	63	56	34
634	Feul & Elec Systems Rep	523	51	61	56	70	22	35	63	71	61	45
635	Track Veh Chassis Rebuild	103	33	48	52	60	42	47	64	62	56	38
710	Clerk	311	62	68	39	43	52	42	41	20	44	30
712	Stenography	569	49	50	38	33	46	32	24	16	28	15
714	Postal Operations	293	65	71	51	48	52	39	47	36	38	26
716	Personnel Admin	286	59	66	43	45	55	40	52	26	45	21
716	Personnel Mgt (Enl)	556	67	70	48	49	53	39	44	20	40	30
717	Adv Army Admin	406	65	63	44	46	53	26	35	34	37	31
753	Machine Acctg	383	60	69	59	50	53	45	39	27	47	33
763	Ord Storage Spec	291	69	75	50	58	57	36	65	45	44	33
911	Medical Aidman, Adv	308	61	61	57	52	44	50	48	34	50	41
912	Medical Tech	271	37	26	17	28	20	16	19	21	26	13
917	Dental Asst	367	51	54	44	41	29	27	38	22	49	37
941	Cook	305	37	39	34	34	32	28	36	33	28	12

Table A-1 (continued)

Current MOS	Course Title	N	ACB Tests RV	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
951	Military Police, Enl Adv	159	61	67	51	52	39	32	49	24	44	38
952	Disciplinary Guard, Enl	144	52	59	41	52	24	32	48	47	45	33
953	Criminal Investigation	192	62	63	51	57	42	24	41	31	54	49
051	Radio Op (Interim Speed)	150	22	18	12	19	22	32	07	13	16	09
052	Radio Op (High Speed)	233	23	27	12	21	23	48	16	30	27	22

*Decimal points omitted in table. These data are from Helm et al. (Note 3).

Table A-2

Validity Coefficients* of ACB Tests for 35 Army School Courses

Group B Samples

Current MOS	Course Title	N	ACB Tests RV	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
271-2	Fixed Sta Radio Rep	323	50	61	55	61	37	47	54	44	71	70
281	Microwave-Mult Channel Rep	236	37	41	44	36	16	21	35	18	58	56
282	Radar Rep	237	34	49	49	47	21	37	47	28	63	56
296	Field Radio Rep	481	41	51	38	50	39	39	46	40	55	55
320	Field Wireman	330	59	66	54	51	53	42	60	42	56	50
403	FC Instrument Rep	214	36	54	55	65	26	35	66	56	56	38
411	Ammunition Supply	702	53	51	43	49	45	25	47	37	27	12
421	Arty Mech-Light Weapons	205	64	65	66	67	60	34	67	64	46	19
422	Small Arms Weapons Mech	430	62	67	62	60	50	38	65	62	46	27
424	Turret Arty Rep	186	61	63	52	59	62	39	61	56	46	24
442	Welder	243	38	52	44	48	32	27	50	45	34	25
443	Machinist	314	50	61	62	55	44	28	62	52	52	40
452	Dental Lab	121	11	26	36	33	-03	18	27	31	17	25

Table A-2 (continued)

Current MOS	Course Title	N	ACB Tests									
			RV	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
631	Automotive Mech	198	57	61	52	66	36	22	60	72	57	49
632	Track Veh Rep	430	65	65	56	67	47	23	72	78	53	16
632	Armor Track Veh Maint	248	45	63	57	59	35	37	60	63	47	24
634	Fuel & Elec Systems Rep	522	55	64	62	69	35	35	69	70	60	35
635	Track Veh Chassis Rebuild	114	50	50	46	56	37	12	51	50	47	30
710	Clerk	340	67	75	56	54	59	43	48	34	52	34
712	Stenography	569	49	51	39	26	49	36	18	03	23	22
714	Postal Operations	295	61	67	53	44	49	40	41	29	38	26
716	Personnel Admin	286	62	65	46	42	45	35	38	21	39	27
716	Personnel Mgt, En'	556	70	69	50	54	54	40	45	19	54	29
717	Adv Army Admin	401	63	67	45	47	53	34	41	22	47	34
753	Machine Acctg	355	60	60	50	49	49	38	45	32	50	33
763	Ord Storage Spec	274	70	75	66	67	60	44	63	50	49	18
911	Medical Aidman, Adv	200	67	67	41	45	45	32	46	17	47	40
912	Medical Tech	200	27	30	25	31	27	26	28	23	21	18
917	Dental Asst	200	65	65	49	48	43	43	36	19	49	41

Table A-2 (continued)

Current MOS	Course Title	N	ACB Tests RV	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
941	Cook	338	28	32	25	26	28	28	30	26	21	06
951	Military Police, Enl Adv	363	60	63	55	51	47	34	51	32	35	23
952	Disciplinary Guard, Enl	139	66	67	44	54	57	39	49	36	43	20
953	Criminal Investigation	106	75	79	67	58	53	60	38	-05	53	51
051	Radio Op (Interim Speed)	145	43	37	34	39	39	59	37	35	16	14
052	Radio Op (Interim Speed)	233	21	11	20	22	19	36	17	11	17	27

*Decimal points omitted in table. These data are from Helm et al. (Note 3).

Table A-3

Tests in the Army Classification Battery

Test	Description
RV	Vocabulary
AR	Arithmetic Reasoning
PA	Pattern Analysis (Spatial Aptitude)
MA	Mechanical Aptitude
ACS	Army Clerical Speed
ARC	Army Radiocode Aptitude
SM	Shop Mechanics
AI	Automotive Information
EI	Electronics Information
RI	Radio Information

Testing for Fairness with a Moderated Multiple Regression Strategy: An Alternative to Differential Analysis

C. J. Bartlett, Philip Bobko, Steven B. Mosier, and Robert L. Hannan

University of Maryland, College Park

Abstract

It is argued that analyses of subgroup differences utilizing a bivariate correlation strategy do not provide an adequate examination of test fairness. An analysis of differential prediction, which involves slopes and intercepts of regression lines, results in more complete coverage of the test fairness issue, since the overall regression line determines the way in which a test is used for prediction. While subgroup correlation coefficients yield information concerning the slopes and intercepts, means and standard deviations must also be examined. A moderated multiple regression strategy is recommended as an alternative to separate analyses by subgroups. An ordered step-up regression procedure is presented which is more encompassing than the bivariate strategies, while avoiding inherent problems associated with subgroup coding in multiple regression.

Testing for Fairness with a Moderated Multiple Regression Strategy: An Alternative to Differential Analysis

C. J. Bartlett, Philip Bobko, Steven B. Mosier, and Robert L. Hannan

University of Maryland, College Park

Single group validity is dead! Differential validity is in intensive care! Although we have come to bury these concepts rather than praise them, please do not forget that the ghost of Caesar lived on. The ghost of single group and differential validity is differential prediction, and it cannot be destroyed merely by wishful thinking or hatred for the EEOC guidelines.

In order for the arguments in this paper to be understood, a clear definition of terms is necessary. Boehm (1972) separated the concepts of single group and differential validity. Unfortunately, differential prediction (Bartlett & O'Leary, 1969) has not been clearly separated from differential validity (e.g., Boehm, 1977). For clarification the three concepts are defined as follows:

(a) single group validity - ". . . where a given predictor exhibits validity significantly different from zero for one group only, and there is no significant difference between the two validity coefficients" (Boehm, 1972, p. 33).

(b) differential validity - "There is a significant difference

between the correlation coefficient of a selection device and a criterion obtained for one ethnic group and the correlation of the same device with the same criterion obtained for the other group" (Boehm, 1972, p. 33). (Boehm added a second restriction that at least one of the validity coefficients be significantly different from zero, and later modified it (1977) to say that both validity coefficients could not be significantly different from zero. This seems to unduly complicate the definition.)

(c) differential prediction - There is a significant difference between the regression equations for two groups as indicated by differences in the slopes, intercepts or both. "Two different groups may differ in test performance, criterion performance, and in how these two relate" (Bartlett & O'Leary, 1969, p. 2). These differences may be reflected in the means, standard deviations, or correlations. Although single group and differential validity consider only the correlation coefficients, Bartlett and O'Leary (1969) urged the consideration of a variety of statistics in interpreting whether a test is discriminatory. Differential prediction considers how a test is used for prediction, which involves a regression equation including means, standard deviations, and correlation coefficients.[1]

---

[1] In the decision of U.S. vs. Jefferson County (1977) the Federal Court noted the difference between differential validity and differential prediction as follows: "Failure, however, to reject the hypothesis that the correlation coefficients are the same for both groups is not by itself sufficient to demonstrate fairness."

It has been asserted that single group validity is dead. As a concept in the study of personnel selection, single group validity has been claimed as merely a chance phenomenon (Humphreys, 1973; Schmidt, Berner, & Hunter, 1973; O'Connor, Wexley, & Alexander, 1975). The final nail in the coffin may have been placed when it was noted: "By definition, the phenomenon does not exist in the population" (Bartlett, Bobko, & Pine, 1977, p. 156). In the population all things are either identical or significantly different. Therefore at a conceptual level, single group validity's existence can never be tested. Certainly, it is never technically feasible to draw any rational conclusions about the population, whatsoever, when the phenomenon of single group validity is observed in a sample.

Differential validity has been similarly claimed as merely a chance phenomenon (Schmidt, Berner, & Hunter, 1973) although the population argument given above does not apply. Some have even gone so far as to say, "Differential validity has been completely discredited as a viable hypotheses. . . . In summary, the theory of black/white differential validity is now adhered to by a small group of vocal professionals, but has been scientifically discredited. . ." (U.S. Chamber of Commerce Adhoc Business Committee on Proposed Employee Selection Guidelines, Note 1). Although this conclusion may be overstating the issue of differential validity, it does represent a widely-held opinion within the industrial/organizational psychology profession. Additionally, an analysis of 1190 comparisons of validity

coefficients from black and white subgroups (O'Leary, Farr, & Bartlett, Note 2; Farr, O'Leary, Pfeifer, Goldstein, & Bartlett, Note 3) showed differential validity in 81 cases (6.81%) at the .05 level. Although this might be interpreted as more than would be expected by chance alone, it demonstrates that differential validity is not a substantial factor for these comparisons.

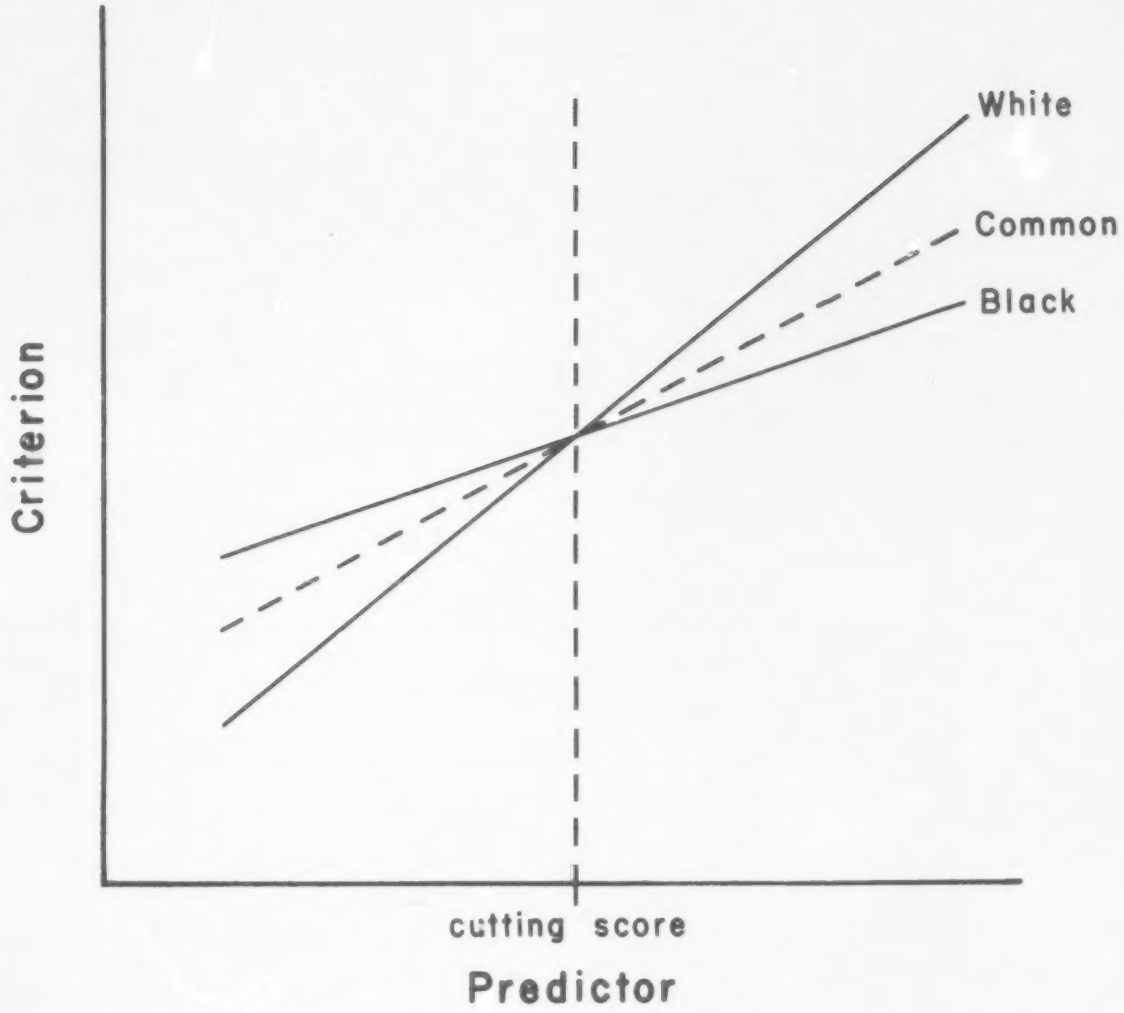
Differential prediction as a substantial phenomenon cannot be written off as easily. The U.S. Chamber of Commerce Adhoc Committee acknowledged, "Finally and perhaps of most importance, it should be noted that the research investigations described above indicated that rather than underpredicting black performance (blacks would do better on the job than test scores would indicate), black performance tended to be overpredicted (blacks do less well on the job than test scores would indicate)" (Note 1). In addition, an examination of the 1190 comparisons (O'Leary et al., Note 2; Farr et al., Note 3) found significant slope differences in 68 cases (5.21%) and significant intercept differences in 214 cases (17.98%). Thus some kind of significant differential prediction was found for 282 (23.19%) of the comparisons. Since these were not independent tests, it is difficult to estimate the exact probability of this occurring, but it is clearly above chance level.

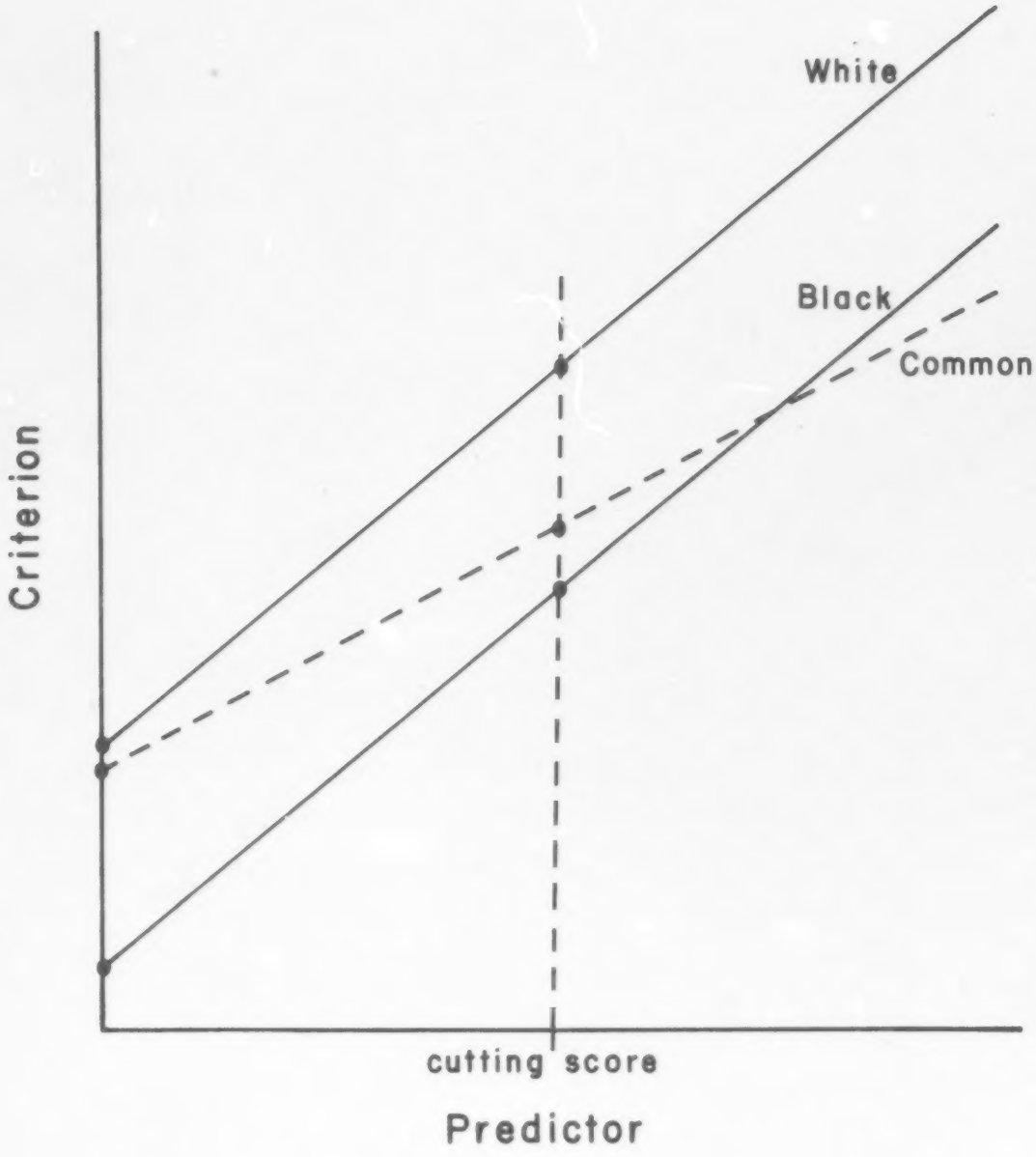
In the cases where significant slope differences occurred, the slopes (and validity coefficients) were more frequently higher for

the white subgroups. If a common regression line were used, this would yield an underprediction for whites and an overprediction for blacks (see Figure 1). Note that to the right of the intersection of the subgroup regression lines, use of a common regression line would result in test unfairness (Cleary, 1968) for whites. To the left of the intersection the reverse would be true (i.e., unfairness for blacks). In this case, where the work group is selected on the basis of the cutting score, a researcher should ultimately conclude that the test is unfair to the white subgroup.

An examination of fairness in instances of differential intercepts would also (probabilistically speaking) lead one to the conclusion that a common regression line would underpredict the performance of whites. This is based upon the finding that the white subgroup exhibits significantly greater criterion mean performance (see Figure 2). Of course these conclusions are based upon the premise that the criterion measure is a valid, nondiscriminatory reflection of performance (a questionable assumption in most situations).

Rather than belabor the potential for overprediction or underprediction of subgroups from data collected prior to the impact of Title VII of the 1964 Civil Rights Act on selection processes, let it be noted that differential prediction can occur, and when it does, the use of a common regression line leads to unfairness. It is a requirement in all of the guidelines that one must consider unfairness. The purpose of the remainder of this paper is to examine alternative


White
Common
Black
Criterion
cutting score
Predictor


White
Black
Common
Criterion
cutting score
Predictor

methods for this consideration.

The Differential Bivariate Analytic Model

The traditional methodology for testing the validity and fairness of a selection measure has been to test the validity for the total sample and then split the sample and test the validities for each of the relevant subgroups. This traditional view claims full support for validity and for test fairness only when (1) the total sample validity is significantly greater than zero, (2) the subgroup validities are both significantly greater than zero, and (3) the subgroup validity coefficients are not significantly different. Success in meeting these stringent requirements is, in large part, a function of sufficient sample sizes for all relevant subgroups (Schmidt, Hunter, & Urry, 1976).

Testing the various validity coefficients of subgroups and the total sample is not the most appropriate methodology because it is not only the correlations that lead to disparate treatment but also the way in which they are utilized in the prediction equation. Thus, tests of the equality of slopes and intercepts should be of primary concern.

Practitioners have argued that tests for differential prediction should not be necessary since (1) subgroup differences do not really exist in the population, (2) test unfairness is typically found for whites rather than for the protected classes and (3) it is rarely

technically feasible to test for differential prediction because of sample size limitations for at least one subgroup (U.S. Chamber of Commerce Adhoc Business Committee on Proposed Selection Guidelines, Note 1). These three arguments not only fail to satisfy the requirements dictated by various guidelines, but they are also contradictory. They say, in essence, (1) differential prediction is a chance phenomenon, (2) differential prediction is not a chance phenomenon but it only removes unfairness toward whites, (3) the true nature of differential prediction is not discernable.

Of course, avoiding a test for differential validity altogether avoids the negative consequences of detecting single group validity (which has no rational explanation as a population concept) or of finding differential validity. However, some kind of test for the effects of heterogeneity is required by several proposed guidelines and subsequent court interpretations. Thus avoidance of the problems and wishful thinking do not solve the problems but function to aggravate and prolong them. This is especially salient in situations where negative impact of a selection device against protected classes exists. As long as there is differential subgroup performance on either the test, the criterion, or both, the possibility of differential prediction <u>is</u> real. For example, consider the regression equation

$$Y = \bar{Y} - \frac{sy}{sx} r_{xy} \bar{X} + \frac{sy}{sx} r_{xy} X$$

If either the means, standard deviations, or the correlations are different for different subgroups, then the resulting equations may also be different for the subgroups in question.

The Moderated Multiple Regression Strategy

An alternative procedure for the determination of validity and differential prediction is available within the rubric of a moderated multiple regression strategy. Such a procedure need not be carried out in conjunction with the separate group bivariate analysis, since moderated multiple regression subsumes the bivariate analyses. Although similar procedures have been suggested previously (Saunders, 1956; Hunter & Schmidt, 1976), the implications of this procedure and the procedure itself have not been explicitly developed in the context of equal employment issues.

For simplicity, consider the situation where one is interested in validating an ability test (A) against a performance criterion (Y) and also in examining the effects of culture (C). The following multiple regression equation accounts for the same information and yields more information than the separate group analysis.

$$(1) \quad Y = b_a A + b_c C + b_{ac} AC + K_1$$

Here, b_a is the regression weight for ability, b_c is the regression weight associated with culture, b_{ac} is the weight for the interaction

term between ability and culture, and K_1 is the additive constant. A significant weight for ability (b_a) indicates significant validity for the test over and above any cultural effect regardless of the significance or non-significance of b_c and b_{ac}. For those who theorize that differential prediction is an artifact properly relegated to chance, their hypothesis is refuted when either or both of the weights associated with culture (i.e., b_c and b_{ac}) are significant. If, on the other hand, non-significant regression weights are found for culture (b_c) and/or the interaction between culture and ability (b_{ac}), no support for the notion of differential prediction has been obtained. A significant weight for culture (b_c) reflects a difference in the intercepts for the two groups that cannot be accounted for as a function of slope differences. A significant ability-culture interaction weight (b_{ac}) indicates differential regression slopes.

However, generation of equation (1) in the sample may result in spurious conclusions as a function of an arbitrary coding scheme (Gocka, 1974). Coding is always arbitrary when assigning numerical values to discrete categorical variables such as race. Any systematic coding scheme will not affect the intercorrelations among ability, culture and performance; however, the correlations between the interaction term and these measures are not independent of the coding scheme used. The overall multiple correlation from equation (1) is independent of the coding scheme, but the profile of the regression weights is not. Therefore, interpretation of individual regression

weights may lead to spurious conclusions.

A procedural strategy has been developed which allows for conclusions that are independent of the coding scheme. This strategy involves an ordered step-up regression procedure. The a priori ordering is consistent with the need to test for validity and for fairness (Cleary, 1968), while accounting for the practical problems of cross validation when differential prediction is found in the back sample.

Suggested Strategy

Step 1. First compute equation (2) for the total sample. This

$$(2) \quad Y = b_a A + K_2$$

determines whether there is a significant overall relationship between ability and performance. If there is not, a number of explanations can be considered. For example, one may simply conclude that the test is not a useful predictor. However, it is also possible that the lack of overall validity is a result of the heterogeneity of the groups with regard to performance as a linear function of ability. If such a case of differential prediction is hypothesized, one can go directly to equation (1) and utilize a procedure such as Wherry Test Selection (Wherry, 1940) to select the combination of terms most likely to be successful for differential prediction. Empirical cross validation of any resulting prediction equation remains a prerequisite to practi-

cal implementation. From the point of view of fairness, when an overall relationship is demonstrated between ability and performance, differential prediction is still a possibility and should be investigated by the procedure suggested in Step 2.

Step 2. Since the most common form of differential prediction appears to be a function of differential intercepts, investigation of this phenomenon is recommended as a next step. Adding the cultural term to the ability term as shown in equation (3) provides a test for

$$(3) \quad Y = b_a A + b_c C + K_3$$

this hypothesis. If culture (b_c) does not add significantly to the prediction, the test can be recommended as fair for use by the Cleary (1968) definition and should be acceptable under any of the guidelines. However, if b_c is significant, Step 3 is recommended.

Step 3. If culture (b_c) adds significantly to the prediction, it is recommended that the product term (b_{ac}) be included, as in equation (1). If the product term adds significantly to the prediction, then equation (1) must be cross validated prior to implementation. If the product term (b_{ac}) does not add significantly, then the required cross validation should be performed using equation (3). The final confirmation of any differential equation requires that the addition of the cultural term(s) in the multiple regression equation results in an increment in predictability beyond that obtained by ability alone. Furthermore, the improvement in prediction must be

maintained in the cross validation sample.

Conclusion

It is always possible to test for differential prediction by any strategy rather than trying to justify the lack of a test on grounds of technical infeasibility. After all, it is always technically feasible to analyze your data. If you do not do it, some one else may do it for you and draw the undesirable conclusions you wished to avoid by claiming that it was not technically feasible to test for differential prediction.

Following the moderated multiple regression strategy allows for the testing of differential prediction with fewer liabilities than a separate group analysis. The need for explanation of single group validity findings is eliminated, since the case of single group validity would be properly relegated to chance. When sample size is not large enough to detect significance, one is left with the conclusion that either differential prediction does not exist, or at least, it is not technically feasible to detect it. Either of these conclusions should satisfy the guidelines.

If differential prediction is detected, a method for dealing with it is provided. Only where differential prediction is detected, and the multiple regression equation does not cross validate to provide improvement, is one placed in the difficult position of suspecting

unfairness, yet not being in a position to remove it by regression analysis.

Finally, with the moderated multiple regression strategy, the moderator need not be confined to discrete groups as in the differential analysis model. Differential prediction between racial or sex groups tells us little in terms of psychological explanations. The more interesting variables from a psychological point of view are likely to be the degree of cultural deprivation, or other situational or personal factors that may explain why subgroup differences exist. Furthermore, such moderators may provide us with more psychological meaning and higher levels of measurement than dichotomies such as race, sex, etc.

Reference Notes

1. U.S. Chamber of Commerce Adhoc Business Committee on Proposed Selection Guidelines. Personal communication, January 19, 1976.

2. O'Leary, B. S., Farr, J. L., & Bartlett, C. J. Ethnic group membership as a moderator of job performance. AIR 753-4/70-TR-1. Silver Spring, Maryland: American Institutes for Research, 1970.

3. Farr, J. L., O'Leary, B. S., Pfeifer, C. M., Goldstein, I. L., & Bartlett, C. J. Ethnic group membership as a moderator in the prediction of job performance: An examination of some less traditional predictors. AIR 73-9/71-TR-2. Silver Spring, Maryland: American Institutes for Research, 1971.

References

Bartlett, C. J., Bobko, P., & Pine, S. M. Single group validity: Fallacy of the facts? Journal of Applied Psychology, 1977, 62, 155-157.

Bartlett, C. J., & O'Leary, B. S. A differential prediction model to moderate the effects of heterogeneous groups in personnel selection and classification. Personnel Psychology, 1969, 22, 1-17.

Boehm, V. R. Differential prediction: A methodological artifact. Journal of Applied Psychology, 1977, 62, 146-154.

Cleary, T. A. Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-123.

Gocka, E. F. Coding for correlation and regression. Educational and Psychological Measurement, 1974, 34, 771-783.

Humphreys, L. G. Statistical definitions of test validity for minority groups: Reality or illusion. Journal of Applied Psychology, 1973, 58, 1-4.

Hunter, J. E., & Schmidt, F. L. Critical analysis of statistical and ethical implications of various definitions of test bias. Psychological Bulletin, 1976, 83, 1053-1071.

O'Connor, E. J., Wexley, K. N., & Alexander, R. A. Single group validity: Fact or fallacy. Journal of Applied Psychology, 1975, 60, 352-355.

Saunders, D. R. Moderator variables in prediction. Educational and Psychological Measurement, 1956, 16, 209-222.

Schmidt, R. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests. Journal of Applied Psychology, 1973, 58, 5-9.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion related validity studies. Journal of Applied Psychology, 1976, 61, 473-485.

United States of America v. Jefferson County. CA 75-P-066-S, 1977.

Wherry, R. J. The Wherry-Doolittle test selection method. In W. H. Stead & C. L. Shartle (Eds.), Occupational counseling techniques. New York: American Book Co., 1940.

Moderators and Subgroups

William A. Owens

The University of Georgia

Abstract

It is argued that an alternative preferable to the use of moderators is subgrouping subjects on the basis of the patterns or profiles of their scores across several dimensions of significance. This means that the variance between subgroups is of primary concern. Evidence is presented to suggest that, as implied, the simple fact of subgroup membership will often significantly enhance both meaning and prediction.

Moderators and Subgroups

William A. Owens

The University of Georgia

As classically defined, a moderator variable is one at different levels of which the relationships between a second and third variables may be expected to differ.

We are all familiar with a more or less common list of the difficulties encountered in moderator variable research. (1) Sample size is reduced. (2) Range of talent is restricted. (3) If multiple R is used, an overall "beta" may cross validate better than the average of n within subgroup "betas." (4) If prediction is enhanced, it may still be at a score level which is not of interest. Thus, aptitude may predict achievement better at high levels of compulsivity, but this fact may be of scant interest to a tutor of non-compulsive athletes. (5) Our moderator variables themselves may lack substantial validity, interpretability or theoretical significance. In a sense, we will also be confounding on all the other significant dimensions which we do not identify and have not measured. (6) In a similar vein, it may be said that single moderator research is not very satisfying because we normally know so little about the "kinds of people" for whom a given device predicts, or fails to predict, a given cri-

terion. Once we begin to think in terms of multiple moderators, we encounter profiles of scores. If some of these profiles are similar, and can be clustered or subgrouped, we have then indeed moved toward the identification of what may be regarded as "kinds of persons."

Let us assume, for the moment, that we are able to subgroup people on a set of reasonably valid and representative predictors. What will be the conceptual sources of variance in scores along any given predictor dimension? First, there will be differences between the means of subgroups; second, there will be within subgroups individual differences; and third. within subgroups error. However, as the subgrouping approaches perfection, the within subgroups individual differences tend to disappear and to leave only error. Thus, intra-subgroup differences in the prediction of, let us say, achievement from aptitude would vary from subgroup to subgroup on a largely random and non-reproducible basis.

Granting a somewhat idealized hypothetical case, the foregoing would seem to argue that the significant variance is <u>between</u> subgroups and that <u>the fact of subgroup membership itself may often be a most significant datum</u>. If this can be documented, we may then ask a final question regarding whether or not subgrouping offers more to the enhancement of prediction than the use of moderators.

What is the evidence that subgroup membership, per se, predicts a broad spectrum of behaviors? A brief explanation of the context and circumstances would seem to be required. At The University of

Georgia we have been administering a scored autobiographical data form to entering freshmen and subgrouping the respondents in terms of the patterns of their response (Owens, 1971). The steps involved are somewhat as follows: (1) Item intercorrelations are factored by the method of principal components (males and females separately); (2) Each subject is assigned a score on each component, thereby generating a profile; (3) A matrix of the distances between each profile and each other is constructed employing the Cronbach and Gleser (1953) D^2 as the metric; (4) The Ward and Hook (1963) hierarchical procedure is applied to this matrix of intersubject distances as a means of assigning the profiles to clusters or subgroups; (5) Because the grouping procedure is hierarchical, the ultimate assignment of subjects to subgroups is reaffirmed in several ways.

In any event, once subjects have been assigned to a cluster, the basic issue can be confronted. We have argued that biodata is a measure of prior experience; and that since past behavior is the best predictor of future behavior, subjects within our subgroups who have behaved similarly in the past should continue to do so in the future. Thus, subjects within a cluster should behave with internal similarity and external differentiation. As regards the former, it may be hastily observed that, on a group of independent reference or marker variable measures, within subgroup sigmas tended to be about 7/10 as large as overall sigmas; and that subgroup means differed as widely as one and one-half units of the overall standard deviation

(Table 1). On the latter, we have now conducted a series of some 42 experimental and field studies to test for the behavioral reality of significant subgroup differences and in approximately 84% of the cases we have found such differences. A typical study from each of seven psychological domains appears in Appendix A. If the differences represented are summarized, subgroup by subgroup, one obtains the sorts of protocols which are illustrated in Table 2. Such data suggest that the unembellished fact of subgroup membership does provide an excellent prediction of behavior.

A critical pair of questions remains. (1) Just how good is this prediction; and (2) does it yield anything, beyond greater generality, not obtained by applying multiple R directly to the biodata factor scores. With respect to goodness of prediction, many illustrations are possible, but a study by Schoenfeldt (Note 1) seems best because it deals with the familiar criterion of academic performance. Table 3 contains some major findings. It will be noted that half the members of one subgroup are gone by the end of their freshmen year; whereas 93% of the members of another subgroup are still present. Moreover, if the subjects of 1970 are "slotted into" the subgroup structure derived in 1968, the mean grade-point averages of the corresponding sets of subgroups are found to be correlated 0.89 ($p<0.001$). This kind of finding clearly demonstrates a substantial and stable differential affinity of subgroup for the criterion.

The second question on incremental validity may be formulated

Table 1

Magnitudes of Differences Between Subgroup Means on Reference Measures (Males)[a]

		Reference Measures																								
Group	N	1	2	3	4	5	6	7	8	9	10	11[b]	12	13	14	15	16	17	18	19	20[b]	21	22	23	24	25	26
1	40																										
2	39	H							H																		
3	62	L	L		L															H				H			
4	31					H	L	L		L				L		L				L			H	L			
5	25															H						H					
6	16						H		H	H				H								H	H		H	H	H
7	12	H	H	H	H											L						L	L			L	H
8	39															H										H	
9	29		L	L	L										H				H	H			H	H			
10	24							H		H													L	H			H
11	20		H		H	H		L		L					L		L				H		H	L		H	
12	39																										
13	54																										
14	26	H													L	L		H									
15	22												H			H											
16	22		H		H		H				H		H							L				L			
17	20	H	H	H	H																	L					
18	19				H						L	H								H						H	
19	10					H	H								L	L			H								L
20	36	L																								H	
21	22	H				L											L					L	L			L	
22	30																									H	
23	21	L				H		L		L	L		L	L		L									L		
Total	658																										

[a]H = Group mean is more than .5 S.D. above grand mean.

L = Group mean is more than .5 S.D. below grand mean.

[b]Differences among subgroup means were significant, excepting variables 11 and 20.

Table 2

A Typical Subgroup Protocol (Non-Conformist Leaders)

Male
Subgroup #4

I. Biodata Inventory Results:

High Factors	Low Factors
1.	1. Religious Activity
2.	2. Athletic Interest
3.	3.
4.	4.
5.	5.

II. Marker Variable Results:

High	Low
1.	1. Tender Minded
2.	2. Direct F
3.	3. Total F
4.	4. Social Religious Conformity
5.	5. Negative Emotionality
	6. Conceptual Simplicity
	7. Neuroticism

Table 2 (continued)

III. Strong Vocational Interest Blank:

High	Low
1. Ad. Man.	1.
2. Psychiatrist	2.
3. Librarian	3.

IV. Field Studies:

		Mean
1. % in Arts and Sciences	55%	(55%)
2. % on Dean's List	12%	(10%)
3. % on Probations	9%	(33%)
4. % of Dropouts	24%	(25%)

This group (male #4) along with male group #23 scored a significantly higher frequency of "good responses" on 5 out of 10 blots on the Harrower, Group Rorschach, than did male groups #1 and #21 (Frazer Study, Note 3).

Wright (Note 8) found this subgroup (low authoritarianism, high complexity) with subgroup 11 performed significantly better ($p \leq .10$) than subgroups 10 and 21 combined in a reading comprehension task in which amount and subject matter read were varied in order to have more (less) complicated tasks.

Table 2 (continued)

Boardman et al. (1974) found this subgroup to contain a significantly high number of leaders.

Schoenfeldt (1974) found this subgroup (N = 23) to differ markedly in the academic majors they chose versus the total sample: science 26.1% (16.6%); speech-journalism 26.1% (9.5%); and social science 4.3% (10.8%).

Strimbu (Note 6) found that percentages of drug users within a given subgroup varied from 0 to 75, thus yielding a highly significant chi-square ($p<0.01$). This subgroup was in the top quartile (high drug usage).

An atypically high number of this group were found to be overachievers (low SAT, high GPA). (Klein Study, Note 9).

Membership in this subgroup is a fairly strong indicator of homosexuality. (Lewis & Schoenfeldt Study, 1973).

Table 3

Subgroup Differences on Educational Criteria

N	Group	% Arts & Sciences School	% Dean's List	% Probations	% Dropouts	Group
44	1	45	5	39	18	1
48	2	69	17	25	Lo- 8	2
73	3	49	Lo- 1	Lo- 9	25	3
33	4	55	12	Lo- 9	24	4
31	5	57	13	35	13	5
17	6	56	12	12	29	6
14	7	Hi-79	21	Lo- 0	Lo- 7	7
48	8	57	4	25	27	8
34	9	47	3	Hi-59	24	9
29	10	55	7	28	34	10
25	11	56	8	20	20	11
42	12	45	19	29	31	12
58	13	Lo-40	3	45	36	13
27	14	Hi-77	15	18	15	14
25	15	52	8	28	32	15
24	16	67	8	21	17	16
22	17	62	Hi-55	Lo- 9	Lo- 9	17
22	18	50	5	45	23	18
14	19	Hi-93	14	21	21	19
45	20	67	4	Hi-53	36	20
15	21	64	Hi-40	12	12	21
37	22	65	3	Hi-54	38	22
24	23	62	Lo- 0	46	Hi-50	23
Average %		55%	10%	33%	25%	

in at least two ways. (a) Does the subgrouping model sometimes succeed when a moderated multiple R fails? A study by Pinto (Note 2) is suggestive. In it a comprehensive biodata form, plus certain standardized tests, were administered to 2,060 salesmen. For purposes of analysis the sample was split into subsamples A and B. The subjects of subsample A were then subgrouped, and the subjects of subsample B were assigned to those subgroups with only a 10 percent loss in fit. Termination reports constituted the criterion, which was predicted via three models. In the first, other measures were employed to predict within biodata subgroup, but no moderator effects were observed. In the second, a multiple R generated in subsample A from the non-biodata predictors vanished on cross-validation in subsample B. However, in the third, the biodata subgroups showed clear differential affinities for the criterion, and the magnitudes of these affinities correlated 0.66 ($p<.01$) in subsamples A and B. (b) A more direct answer to this same question may be evoked by asking does knowledge of subgroup membership enhance prediction <u>beyond</u> what could be obtained from the basic factor scores alone. A study by Feild (1975) is germane. He evaluated data obtained from 509 college students who had completed a biodata form as entering freshmen and a so-called College Experience Inventory (C.E.I.) as graduating seniors. Criteria for prediction were 24 factor scores (12 for each sex) derived from the C.E.I., and canonical correlation was employed to relate biodata factor scores plus dummy coded subgroup information

to each of these criteria. It was found that in four of 24 instances, subgroup information significantly enhanced predictions based upon the biodata factor scores alone.

It has been the purpose of this discussion to suggest that a knowledge of (biodata) subgroup membership which capitalizes on between-subgroups variation, may sometimes be more interesting and contribute more to prediction, than a moderator approach which depends upon the presence of non-random variance within groups. In addition, subgroup means are quite stable, and subgroup profiles across several dimensions are often highly interpretable.

Reference Notes

1. Schoenfeldt, L. F. Life experience subgroups as moderators in the prediction of educational criteria. Paper read at the meeting of the American Educational Research Association, Minneapolis, 1970.

2. Pinto, P. R. Subgrouping in prediction: A comparison of moderator and actuarial approaches. Unpublished doctoral dissertation, University of Georgia, 1970.

3. Frazer, R. W. Differential perception of individuals subgrouped on the basis of biodata responses. Unpublished doctoral dissertation, University of Georgia, 1971.

4. Helms, W. Biodata subgroup differences in recall and clustering of interest area stimulus words. Unpublished M.S. thesis, University of Georgia, 1972.

5. Brush, D. H. Predicting major field of college concentration with biographical and vocational interest data: A longitudinal study. Unpublished M.S. thesis, University of Georgia, 1974.

6. Strimbu, J. J. A quasi-actuarial approach to the identification of college student drug users. Unpublished doctoral dissertation, University of Georgia, 1973.

7. Piacentini, J. J. Physical correlates of prior behavior pattern. Unpublished M.S. thesis, University of Georgia, 1974.

8. Wright, R. B. Some biodata subgroup differences in reading comprehension. Unpublished M.S. thesis, University of Georgia, 1973.

9. Klein, H. A. Personality characteristics of discrepant academic achievers. Unpublished doctoral dissertation, University of Georgia, 1972.

References

Boardman, W. K., Calhoun, L. G., & Schiel, J. H. Life experience patterns and the development of college leadership roles. Psychological Reports, 1974, 31, 333-334.

Cronbach, L. J., & Gleser, G. Assessing similarity between profiles. Psychological Bulletin, 1953, 50, 456-473.

Eberhard, C., & Owens, W. A. Word association as a function of bio-data subgrouping. Developmental Psychology, 1975, 11, 159-164.

Feild, H. S. The utility of homogeneous subgroups and individual information in prediction. Multivariate Behavioral Research, 1975, 10, 449-462.

Lewis, M. A., & Schoenfeldt, L. F. Development-interest factors associated with homosexuality. Journal of Consulting and Clinical Psychology, 1973, 41, 291-293.

Owens, W. A. A quasi-actuarial basis for individual assessment. American Psychologist, 1971, 26, 992-999.

Schoenfeldt, L. F. The utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with "jobs." Journal of Applied Psychology, 1974, 59, 583-595.

Ward, J. H., & Hook, M. E. Application of an hierarchical grouping procedure to a problem of grouping profiles. Educational and Psychological Measurement, 1976, 23, 69-81.

Appendix A

Typical Studies of Subgroup Differences in Seven Psychological Domains

Perception

Frazer (Note 3) administered the Harrower, Group Rorschach, to 2 pairs of subgroups differing widely on their overall biodata profiles. He found significant subset differences in the frequency of "good responses" on 5 of 10 blots (a sixth was marginal). On 3 of the remaining 4, the "common response" was so common as to virtually preclude the discovery of differences. It is implied that differing norms should exist for subjects of differing background.

Cognition, Creativity, Decision-Making

Eberhard and Owens (1975) related background, as measured via biodata subgroup, to several measures of verbal association derived from the Kent-Rosanoff word list. Four contrasting female subgroups were selected in terms of hypotheses derived from the research literature and sampled systematically to obtain a total of 110 subjects. Word associations were scored on the basis of the protocols of commonality, mean reaction time, object referent response set, and concept referent response set. Analyses were then made to discover whether or not the subgroups could be discriminated in accordance with expectation by scores on each of the six protocols, and six significant outcomes were obtained.

Learning and Memory

Helms (Note 4) put the members of six preselected male subgroups through a learning and recall experiment in which it was hypothesized that a subset with higher measured interest in a given area would better recall and cluster stimulus words relevant to the area than a paired subset with lower interest. In a second phase of the study, the meaningfulness of the stimulus words was evaluated via free association. Results showed significant differences in recall and clustering over trials, and a significant subgroup difference in the mean number of words recalled during trial one by the contrasting high and low interest subsets. There were, however, no significant differences between subgroups in amount of clustering or in number of word associations generated, although all observed differences were in the hypothesized direction.

Interests, Attitudes, Values and Motives

Brush (Note 5) has completed a searching analysis of biodata scores vs. scores on the Strong V.I.B. as predictors of less specific curricular choice and of more specific, vocationally-related choice. In brief, biodata predicts the former criterion as well as the Strong, but the Strong is a better predictor of the latter.

Personality

Strimbu (Note 6) administered a questionnaire containing 30 response options pertaining to drug usage to 425 undergraduate males

at the University of Georgia. He obtained a 42.3% return, and distributed the questionnaires to the 23 biodata subgroups to which his subjects had been priorly assigned. Percentages of users within a given subgroup varied from 0 to 75, and discrepancies between expected and obtained values yielded a highly significant chi-square ($p<0.01$).

Social Processes

Boardman et al. (1974) investigated the hypothesis that college leadership has affinity for, and can be predicted from, biodata subgroup membership. Subjects were an 80% sample of the freshman class of 1968 at The University of Georgia; i.e., 1037 males and 897 females who completed the biodata form during the summer or early fall and were cast into subsets immediately thereafter. During what should have been their Junior year ('70-'71), when 572 males and 535 females remained enrolled, number of offices held in 253 University-recognized campus organizations was determined by subgroup. For males, but not for females, a χ^2 test applied to observed vs. expected values was significant at $p<.01$. Indeed subgroups which contained only 22% of the subjects contained over 59% of the leaders and differed significantly from all other subgroups.

Physical or Physiological

Piacentini's (Note 7) hypothesis was that subjects cast into subsets homogeneous for pattern of prior experience would differ with respect to one or more of three physical indices. The indices selected

were weight, height/weight ratio, and weight differential from freshman to senior years. Six one-way ANOVAS were run, three for males ($n = 300$) and three for females ($n = 236$) to determine the significance of subgroup differences by physical index. For males weight differences were significant at $p<.02$; but for females none were significant. The Newman-Keuls procedure revealed no significant pairwise differences between the ordered mean weights of male subgroups, nor did the Scheffé test reveal any significant weight differences between the means of logically contrasting subsets. Finally, rank order correlations of subgroup means on the physical indices with subgroup means on the biodata dimensions produced seven correlations which would have been significant if postulated in advance but which were not significant, post hoc.

Realistic Job Previews:
Can a Procedure to Reduce Turnover Also Influence the Relationship between Abilities and Performance?

John P. Wanous

Graduate School of Business Administration

Michigan State University

Abstract

The impact of realistic job previews (RJP's) as a moderator of the ability-performance relationship is evaluated from two perspectives. First, the research evidence from six experimental studies indicates that RJP's have no impact on either the level of job performance or the ability-performance relationship. The primary reason for this is that the RJP concerns the matching of human needs to organizational climate, and thus is designed to influence job satisfaction and turnover rather than job performance. Second, the potential impact of RJP's on the ability-performance relationship is the subject of speculation in three areas: self-selection, ambiguity reduction, and the timing of the RJP during organizational entry.

Realistic Job Previews: Can a Procedure to Reduce Turnover Also Influence the Relationship between Abilities and Performance?

John P. Wanous
Graduate School of Business Administration
Michigan State University

The question posed in the subtitle captures the theme of this paper. The current state of theory, research, and practice in realistic job previews (RJP's) for recruitment is still immature. To date, the primary focus of these efforts has been to reduce employee turnover, but not necessarily to influence the subsequent job performance of newcomers. Because the empirical evidence of RJP's on performance and the ability-performance relationship is scanty, the bulk of this paper will concern the background, the research results, and unanswered issues concerning RJP's.

What is an RJP?

Background

Many organizations prefer to have a low selection ratio at the entry level. This stems from an overemphasis on "being selective," and it does increase the "utility" of a selection procedure to have

a low selection ratio. Let's call this the traditional approach to recruitment (Wanous, 1975) or the "flypaper" theory (Schneider, 1976) of attraction and retention.

Those who advocate the RJP, such as myself, tend to take a broader view of recruitment. For one thing, recruitment and retention are viewed as closely linked together, rather than as separate processes. A second difference is that two types of "matches" (or correspondences) are recognized as important during the entry of newcomers (Wanous, in press). One of these concerns the matching of abilities and organizational ability requirements. The other is the degree of correspondence between an individual's needs and the reinforcing properties of the organizational climate. The former match-up affects job performance more than turnover; the influence of the latter is the reverse. It should be noted that this distinction has been with us for some time; it is traceable back to March and Simon (1958) and to the Minnesota Studies on Work Adjustment (Lofquist & Dawis, 1969).

Expected Impact of an RJP

Given the conceptual models (Lofquist & Dawis, 1969; March & Simon, 1958) already mentioned, it is a rather straightforward issue to predict that an RJP will have its greatest impact on the correspondence between human needs and the organizational climate, and thus on turnover rather than performance. If this were all there is to the analysis, the paper could end here. A more detailed assessment of RJP theory and practice, however, will reveal the

potential for a limited impact on performance levels and a possible strengthening of the ability-performance relationship.

Besides the expected impact on turnover, there is the theoretical issue of how an RJP acts to reduce turnover. Thus far three reasonable explanations have been proposed, but none has received experimental support to the exclusion of the others. First, realism may reduce turnover by the direct effect of "innoculating" (McGuire, 1964) newcomers against unpleasant realities. This explanation focuses on the matching of expectations and reality--a cognitive theory of turnover. This also seems reasonable, since many theories of job satisfaction use the discrepancy approach (Wanous & Lawler, 1972). In summary, the logic goes that expectations are matched to experience, which reduces the gap between expectations and reality, and therefore increases satisfaction. Since there is a weak but consistent relationship between satisfaction and turnover, the RJP reduces turnover via this cognitive, "innoculation effect."

The second explanation is that people act on their expectations so as to match their own needs to the organizational climate. Thus, an RJP can influence the self-selection (i.e., organizational choice) of newcomers. If this view is correct, then different types of people enter the organization. To a certain extent this explanation depends on the validity of the innoculation effect. That is, the RJP must change expectations (innoculation effect), so that the self-selection effect follows and is contingent upon innoculation.

The third explanation appears to be contingent on the first two. If expectations are innoculated, and if people do make informed self-selection decisions, then they may also feel a sense of freedom of choice during this decision. If this occurs, people will be likely to attribute the locus of the decision to themselves and become committed to it, which reduces turnover.

Impact of the RJP: Research Results

Over the last 20 years a total of only six RJP experimental studies have been conducted (Wanous, in press). The characteristics of each one are summarized in Table 1, which shows the sample size, sex of subjects, job types, the types of RJP used, and when the RJP was presented to subjects during the entry process. The results of each study are shown in Table 2, where they are sorted into: (a) pre-entry (ability to recruit newcomers), (b) entry (influence on initial expectations and influence on choice of the organization by the individual), and (c) post-entry (effects on attitudes, performance, and length of job survival).

The only dependent variable used in all six studies was job survival, and the results confirm the consistent effect of RJP's on this. Only three studies reported post-entry performance levels for the experimental and control groups. None of these found any significant differences between them, although one study obtained results "close" ($p < .10$) to statistical significance.

Table 1

Characteristics of Realistic Job Preview Experiments

	Sample			Experimental Procedures		
Study	Size	Sex	Job Type	Basis for Realism	Means Used	Timing of Preview
Farr, O'Leary, & Bartlett (1973)	N=160	Female	Sewing machine operators at Manhattan Industries, Inc.	2 hr. simulated work experience on sewing machine.	N=80 S's who were tested & did the simulation. N=40 who were tested only. N=40 who neither tested nor did the simulation.	Prior to formal job offer acceptance.
Ilgen & Seely (1974)	N=468	Male	Cadets at the U.S. Military Academy at West Point.	Revision of the Macedonia (Note 1) booklet based on interviews with cadets & officers.	Booklet distributed by mail to N=234 were selected at random as the control.	After written acceptance of appointments, but prior to entry & oath at the 2 mo. summer training program.
Macedonia (Note 1)	N=1260	Male	Cadets at the U.S. Military Academy at West Point.	Questionnaire survey of freshman on time usage & of seniors on perceived climate.	Booklet distributed by mail to N=568 cadets & not to the remaining N=692.	Same as above.

Table 1 (continued)

	Sample			Experimental Procedures		
Study	Size	Sex	Job Type	Basis for Realism	Means Used	Timing of Preview
Wanous (1973)	N=80	Female	Telephone operators at Southern New England Telephone.	Questionnaire survey of experienced operators, interviews with operators & supervisors, & personal observation.	Two films used: the company's recruiting film & a new realistic film both 15 min. long.	After job offer, but prior to formal acceptance of it by the S's.
Weitz (1956)	N=474	Male	Life insurance agents at Life & Casualty Insurance Co., of Tennessee.	Questionnaire survey of experienced agents.	Booklet mailed to prospective agents.	Prior to the organization's selection decision.
Youngberg (Note 2)	N=404	Male	Life insurance agents at Prudential.	Not specified, but said to be similar to the one used earlier by Weitz.	Booklet mailed to applicants.	Prior to organization's selection decision, but some S's had early orientation training as well as the booklet.

Table 2

Results of Realistic Job Preview Experiments

	Pre-Entry	Entry			Post-Entry	
Study	Ability to Recruit	Initial Expectations	Choice of Org. by Person	Attitudes	Performance	Job Survival
Farr, O'Leary, & Bartlett (1973)	Not measured	Not measured	Slightly higher refusal rates for E group (n.s.) & for black vs. white S's (n.s.).	Not measured	Not measured	88.9% vs. 60% ($p < .05$) for 6 wks. & 88.9% vs. 68.8 ($p < .11$) for 4 wks., for white S's. No differences found for black S's.
Ilgen & Seely (1974)	Not measured	Not measured	4.5% of E group withdrew prior to 2 mo. summer training program but no rate for C S's reported.	Not measured	Not measured	94% vs. 88.5% ($p < .05$) for 2 mos.
Macedonia (Note 1)	Not measured	Not measured	10.6% of E group declined vs. 21.1% for C ($p < .01$).	Not measured	Peer ratings of experimental S's lower than for others ($p < .10$).	91.3% vs. 86.1% ($p < .01$) for 1 yr.

Table 2 (continued)

	Pre-Entry	Entry		Post-Entry		
Study	Ability to Recruit	Initial Expectations	Choice of Org. by Person	Attitudes	Performance	Job Survival
Wanous (1973)	No difference between groups.	Lower for E group than for C ($p<.05$ or better). No difference for those facets not in the films.	No difference between groups.	Thoughts of quitting lower for E than C ($p<.05$) after 1 mo.	No difference groups.	62.2% vs. 50% (n.s.) for 3 mos.
Weitz (1956)	No difference between groups.	Not measured	Not measured	Not measured	Not measured	81% vs. 73% ($p<.05$) for all time periods together. 68% vs. 53% for 6 mos. (n.s.).
Youngberg (Note 2)	9% more agents hired in E group (n.s.)	Slightly more realistic for E than C (n.s.).	Not measured	80% vs. 64% "satisfied" at the end of 3 mos. ($p<.001$).	No difference between groups.	89% vs. 84% ($p<.10$) for 3 mos. 71% vs. 57% ($p<.01$) for 6 mos.

Note: E means experimental, C means control.

None of the studies reported the actual correlation between abilities and job performance. In fact, none of the studies obtained ability measures, so that computing a correlation was impossible. Thus, there are no data I know of that directly address the issue of RJP's as moderators of the relationship between abilities and performance.

Besides these six studies of previews, there is another experimental study of realism for newcomers (i.e., post-entry) to reduce the intial anxiety typically associated with entry into a new job situation (Gomersall & Myers, 1966). Assemblers of electronic components at Texas Instruments usually took about four months "to attain a level of competence," which was defined as 85% of the standard rate for the job. Much of this delay in the learning process was attributed to anxiety associated with being a newcomer. To compensate for this an initial group of ten was given a different first day orientation, designed to "soften" the shock of entry. A control group (of unspecified size) of other new hires was inducted in the usual manner, i.e., a two hour session and then put right to work.

After initial success with the small group of ten, the procedure was extended to other groups so that the total of those subject to the new procedure was 100. The authors were careful to point out that the ultimate levels of performance reached were not different between the groups, however, the "realistic socialization" groups attained that level faster. Total cost savings for the 100 experiment-

al new hires were estimated at $50,000 for the first year in job performance, and another $35,000, in the personnel costs of reduced turnover and absenteeism.

How an RJP Could Affect the Ability-Performance Relationship

Via Self-Selection

One of the explanations why the RJP reduces turnover is through its effects on self selection, which facilitates the matching of individual needs to organizational climate. The issue here is just what is being matched up and whether or not perfect compatibility between the individual and the organization is a desirable situation.

Edgar Schein (1968) has speculated that organizational stagnation may occur when newcomers are too closely matched to the climate they enter. He argues that there is a "U-shaped" relationship between innovation and vitality in organizations and the degree to which newcomers conform to the climate. A little bit of deviance is desirable, especially if it occurs in the "peripheral" not "pivotal" values of the organization. Moderately deviant newcomers are seen as "creative individualists" not as "rebels" nor as "conformists."

This logic seems to say that the RJP may lead to conformity and possible stagnation if carried to an extreme. Whether this is so or not depends on the types of personality traits likely to be employed in making the self selection decision. Whether creativity will or will not not be affected by an RJP is unknown at present, but seems

to be a testable proposition.

A related issue is whether or not the RJP can influence self-selection decisions. At the present time there is no conclusive evidence that the preview experience is sufficiently potent to influence this decision. If the RJP turns out to operate mainly through innoculation, then its potential impact on the ability-performance relationships will be lessened.

Via Role Ambiguity Reduction

An expectancy theory of motivation/performance proposed by Porter and Lawler (1968) includes three interactive influences on job performance: abilities, effort, and role perceptions. The purpose of this paper is to focus narrowly on the relationship of abilities to performance, while "controlling" for the effects of the other two. In actual practice, this has been a formidable task for researchers. Role misperceptions can be considered as a source of error variance in performance, while we focus on abilities. In this case, using the RJP, or the Texas Instruments realism treatment for newcomers, may reduce ambiguity, thereby making it easier to detect the "true" relationship between abilities and performance. In cases where some newcomers received a realistic treatment of some sort, and others did not, there could easily be a differences between the correlations obtained in the two groups. A type of "RJP moderator variable" could be the result.

To be an effective "role ambiguity reducer," the RJP should focus on job as well as organizational factors. In the typical RJP both specific (job-related) and general (climate-related) information is included. To date, no study has specifically examined the possible differential impact of one type of information vs. another type. One hypothesis might be that the job component is related to performance, while the organizational component is related to withdrawal from the organization. A climate study by Schneider (1973) suggests that bank customers withdraw their accounts based on broad, climate perceptions, rather than specific factors associated with the particular branch location.

Via Different "Timing"

With one exception, the experiments on realism have been designed to occur prior to organizational entry, and have been aimed at the need-climate relationship. Perhaps a post-entry program of realistic organizational socialization is the more appropriate vehicle for influencing the ability-performance relationship.

Reference Notes

1. Macedonia, R. M. Expectations--press and survival. Unpublished doctoral dissertation, New York University, 1969.

2. Youngberg, C. F. An experimental study of job satisfaction and turnover in relation to job expectations and self expectations. Unpublished doctoral dissertation, New York University, 1963.

References

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. Effect of a work sample test upon self-selection and turnover of job applicants. Journal of Applied Psychology, 1973, 58, 283-285.

Gomersall, E. G., & Myers, M. S. Breakthrough in on-the-job training. Harvard Business Review, 1966, 44, 62-72.

Ilgen, D. W., & Seely, W. Realistic expectations as an aid in reducing voluntary resignations. Journal of Applied Psychology, 1974, 59, 452-455.

Lofquist, L. H., & Dawis, R. V. Adjustment to work. New York: Appleton-Century-Crofts, 1969.

March, J. G., & Simon, H. A. Organizations. New York: Wiley, 1958.

McGuire, W. J. Inducing resistance to persuasion. In, L. Berkowitz (Ed.), Advances in experimental social psychology, Vol. 3. New York: Academic Press, 1964.

Porter, L. W., & Lawler, E. E. Managerial attitudes and performance. Homewood, Ill.: Irwin, 1968.

Schein, E. H. Organizational socialization and the profession of management. Industrial Management Review, 1968, 9, 1-16.

Schneider, B. The perception of organizational climate: The customer's view. Journal of Applied Psychology, 1973, 57, 248-256.

Schneider, B. Staffing organizations. Pacific Palisades, Calif.: Goodyear, 1976.

Wanous, J. P. Effects of a realistic job preview on job acceptance, job attitudes, and job survival. Journal of Applied Psychology, 1973, 58, 327-332.

Wanous, J. P. A job preview makes recruiting more effective. Harvard Business Review, 1975, 53, 166-168.

Wanous, J. P. Organizational entry: Newcomers moving from outside to inside. Psychological Bulletin, 1977, in press.

Wanous, J. P., & Lawler, E. E. Measurement and meaning of job satisfaction. Journal of Applied Psychology, 1972, 56, 95-105.

Weitz, J. Job expectancy and survival. Journal of Applied Psychology, 1956, 40, 245-247.

The Effects of Sex Role Stereotyping on the Ability-Performance Relationship: Prior Research and New Directions

Virginia E. Schein

The Wharton School, University of Pennsylvania

Abstract

Although there is a reasonable body of research pointing up the negative impact of sex role stereotypical thinking on selection decisions, limited attention has been paid to the impact of such thinking on the perceived and actual performance of women in management. A suggested new research avenue is the ways in which sex role stereotypical thinking impacts on organizational factors, such as through differential placement, tokenism and supervisory bias, so as to impair the on the job performance of women managers. Furthermore, it is argued that the relationship between power and political behavior and effective managerial performance needs examination, with particular emphasis on the way in which sex role stereotypical thinking may limit a woman managers' ability to acquire or utilize work related power acquisition behaviors.

The Effects of Sex Role Stereotyping on the Ability-Performance Relationship: Prior Research and New Directions

Virginia E. Schein

The Wharton School, University of Pennsylvania

Sex role stereotyping refers to the belief that a set of traits and abilities is more likely to be found among one sex than the other. The existence of sex role stereotypes has been documented by numerous researchers (Anastasi & Foley, 1949; Maccoby, 1966; Wylie, 1971). For example, Rosenkrantz, Vogel, Bee, Broverman, & Broverman (1968) found that college students perceived men as more aggressive and independent than women, whereas women were seen as more tactful, gentle, and quiet than men. That sex role stereotypical thinking exists among managers within organizations was documented in Schein's 1973 study using 300 middle-line male managers in 12 insurance companies. She found a clear difference between the particular characteristics, traits, and attributes that the respondees perceived to be commonly held by men in general and those the respondees believed to be commonly held by women in general.

More important than the demonstration of sex role stereotyping per se, Schein's study also documented the potential of sex role stereotyping to impact on the ability-performance relationship. Within both the sample of 300 middle-line male managers, as well as

a sample of 167 middle-line female managers (Schein, 1975), she found a strong relationship between sex role stereotypes and the perceived requisite personal characteristics for the middle-management position. According to the results, successful middle managers were perceived to possess those characteristics, attitudes, and temperaments more commonly ascribed to men in general than to women in general. In other words, for both male and female respondees, to "think manager" meant to "think male." With regard to the specific characteristics, Schein found, for example, that both managers and men were perceived to possess the characteristics of leadership ability, competitiveness, self-confidence, objectivity, aggressiveness, forcefulness, being ambitious, and desirous of responsibility. Women were perceived as not possessing these characteristics.

Overall, these results suggest that, all else being equal, the perceived similarity between the characteristics of successful middle managers and men in general increases the likelihood of a male rather than a female being selected for or promoted to a managerial position. As such, this association between sex role stereotypes and perceptions of requisite management characteristics would seem to account, in part, for the limited number of women in managerial positions.

Following these initial studies on sex role stereotyping, other researchers have documented this negative impact of sex role stereotypical thinking on selection decisions. For example, Rosen and

Jerdee (1974) found that respondees, when asked to make managerial selection decisions based on descriptions of applicants who differed only on the basis of sex, tended to make selection decisions in favor of males. Cohen and Bunker (1975), using a similar research technique, found that males, compared to females, were more likely to be selected to a male-oriented position; whereas, females rather than males were more likely to be selected for a female-oriented position. Overall, these and other studies have indicated that sex role stereotyping has a definite and negative impact on the selection of women into managerial positions.

If sex role stereotypical thinking impacts on the perceived potential ability of a woman to perform effectively in an organization, then it would also seem to follow that this same thinking would impact on performance aspects. Although the effect of sex role stereotypical thinking on selection decisions has received a great deal of research attention, such has not been the case with regard to the performance side of the ability-performance relationship. Limited attention has been given to the impact of sex role stereotypes on perceptions of performance and no research efforts have been directed at the impact of sex role stereotypes on on-the-job performance. As such, the purpose of the paper will be two-fold: (1) to examine prior research in the area of performance perceptions and (2) to present new research directions in the area of on-the-job performance.

Perceptions of Performance

Performance Evaluation

One of the key areas that would seem to be impacted by sex role stereotypical thinking is that of performance evaluation. Indeed, two recent studies in this area did find that performance evaluation was influenced by the sex of the performer--in favor of the female. Both Hamner, Kim, Baird, and Bigoness (1974) and Bigoness (1976) found that in situations where performance criteria were objectively defined, high performing females were rated more favorably than high performing males. Given that both studies used "masculine" positions (stock clerks in a grocery store), one explanation of these results is that women in such positions are not expected to do well, based on stereotypical thinking, and when they do perform well they are evaluated more favorably than their male counterparts. In accordance with this explanation, Brief and Wallace (1976), using the less masculine position of library administrator, found that males and females were evaluated and rewarded equally by the respondents.

These studies, then, suggest that in situations in which women, based upon stereotypical expectations, are not expected to do well, but do do well, that they are evaluated more favorably than males who perform in a similar fashion. If so, then we have a situation wherein on the one hand the large body of selection research indicates that women with qualifications equal to men are less likely to be selected for or promoted into managerial positions, and, on the other hand, when they do attain these positions and perform

well, they are likely to receive more favorable performance evaluations than their male counterparts. This explanation, however, seems in direct contradiction to the promotion and progress of women currently in organizations. There has been no evidence to date to indicate that women have been promoted faster or have been perceived as more successful than their male counterparts in managerial positions. Rather, women still seem less likely than their male counterparts to be promoted within the organization.

A recent study by Garland and Price (1977) suggests an alternative explanation to these findings, one which offers a tie-in between research results and the reality of the situation of women in management. In their study, male subjects read descriptions of a successful or unsuccessful female manager and then made causal attributions for her success or failure. In addition, attitudes toward women in management among the respondees were measured with their Women as Managers Scale. They found high correlations between attitudes toward women and the causal attributions of her success. Specifically, they found that males who had negative attitudes toward women as managers were more likely to attribute her success to luck or the ease of the job; whereas, males who had positive attitudes toward women were more likely to attribute her success to her ability or hard work.

These results suggest, then, that although successful women may be perceived even more favorably than successful men, given that this success is contrary to one's expectations, their success

is not necessarily attributed to factors such as ability or doing a difficult job. Hence, on the one hand, stereotypical expectations may produce a positive bias in favor of women with regard to overall evaluation, at the same time stereotypical expectations would attribute this success to factors unrelated to ability. If so, then, women would still not be the recipient of further promotions or better jobs within the organization.

Research merging the approaches of Hamner et al., with those of Garland and Price might provide some interesting insights into the effect of sex role stereotypical thinking on performance evaluations. In addition, research is needed which separates out global performance evaluations from that of evaluations of specific behaviors. While it is still unknown to what extent and how stereotypical thinking influences evaluations within organizations, also unknown is the effect of this thinking on the actual evaluations of the behaviors of women. For example, unknown is the extent to which women who achieve success using stereotypical masculine behaviors, such as aggressiveness, are viewed more or less favorably than women who use stereotypical feminine behaviors, such as intuitiveness and sensitivity, in order to succeed.

Subordinate Perceptions

While no research has been done on the influence of stereotypical thinking on actual behaviors exhibited by women in managerial positions, some research has been done with regard to the evaluation by subordinates of male versus female supervisory behaviors.

For the most part, the thrust of these efforts has been to test a sex role congruency hypothesis, such that female leaders exhibiting behaviors congruent with sex role stereotypes would be evaluated more favorably than males; whereas female leaders exhibiting behaviors not in congruence with sex role stereotypes would be evaluated less favorably than males. Although a variety of researchers have had hypotheses along these lines, the results of these studies have been conflicting and inconsistent. For example, Hansen (Note 1) found no sex of supervisor differences on ratings of support behaviors in facilitation behaviors, although subordinates of both sexes were less satisfied if their supervisor was female. Petty and Lee (1975) conducted a similar study with dissimilar results. Ratings of leader consideration were more highly correlated with subordinate satisfaction when the supervisor was female. There was also a negative correlation between initiation of structure and subordinate satisfaction for a subset of males under female supervisors. These results were interpreted as supporting the sex role congruency notion since the display of consideration behaviors is consistent with the feminine stereotype while the display of initiation of structure behaviors is not. Rosen and Jerdee (1973) hypothesized that feminine behaviors, such as consideration, would be viewed favorably when performed by women supervisors; whereas masculine behaviors exhibited by women would be evaluated negatively. Their results only partially supported their hypotheses. To date, this stream of research has provided very little in the way of consistent information.

Overall, prior research with regard to the impact of sex role stereotyping on perceptions of performance has produced some interesting ideas, yet very little definite outcomes. In part the latter appears due to unsystematic research efforts and in part due to the complexity of the issue. If the impact of sex role stereotyping on the ability-performance relationship is to be examined seriously, there is no doubt that the vital component of performance evaluation deserves high priority research consideration.

On-The-Job Performance

While research efforts directed at the impact of sex role stereotyping on perceptions of performance of women is important, even more important is the need for future research on the impact of sex role stereotyping on the actual performance of women in managerial positions. In what ways can and does stereotypical thinking impact on women's ability to actually function effectively in the managerial role?

Placement

There are some obvious, albeit still unresearched, answers to this question. First of all, stereotypical thinking at the point of entry into the organization can produce differential placement of males and females. For example, if women are seen as more likely to possess characteristics such as humanitarian values, helpful, aware of feelings of others, etc., they would be more likely to be

placed in staff positions, as opposed to line positions. Once in these positions women would be less likely to acquire managerial/ administrative skills and be less likely to be promoted into the more upwardly mobile line functions.

Tokenism

A second way in which sex role stereotyping can influence a woman's ability to perform well is through the placement of a woman into a job based on her sex as opposed to her abilities. On-going affirmative action pressures have forced companies into placing more women into middle and upper level managerial positions. However, to the extent that individuals making these promotional decisions still feel that women are less qualified than men to be managers, they are more likely not to place women into these positions on the basis of their possessing the necessary skills and abilities, but rather to place them simply to get a woman on the job. If this occurs, and the necessary training for this position does not follow, then there is a high probability that a woman would not perform effectively on this job. If so, then the stereotype that women do not make good managers is simply reinforced in the organization and perpetuates either the underutilization of women or continued "token" placements.

Supervisory Bias

A third factor which can inhibit a woman's ability to perform well on the job may be stereotypical thinking on the part of her

superiors with regard to her ability to function effectively as a manager. To the extent that a woman's supervisor believes that she is less likely than a male to be aggressive, forceful, competitive, or ambitious, he or she may be less likely to provide her job assignments in which these skills and abilities appear to be necessary. If so, these differential task assignments can prevent her from learning or developing certain essential administrative abilities and/or produce within the organization an image that she cannot perform these tasks, hence limiting her future promotional progress. Furthermore, to the extent that stereotypical thinking within a supervisor becomes negative bias toward a woman, he or she may limit the amount of information that is given to the female subordinate, exclude her from certain meetings, and provide negative and demotivating direct or indirect cues with regard to her managerial potential.

Power and Political Behaviors

While factors such as differential placement, token promotions, and superior bias, among others, have yet to be given serious research consideration, there is another factor which would seem to have a major impact on a woman's ability to function effectively as a manager--that of exclusion from the power and political networks within the organization, and limited ability to develop power acquisition behaviors. Not only has the role of power and political behaviors with regard to the effective functioning of women managers been neglected, but the entire area until recently has been given limited attention within the field of industrial-organizational

psychology and the more broad field of organizational behavior in general. Despite this lack of research, it is argued here that not only are the development and use of power acquisition behaviors important to effective managerial functioning, but also that a woman's lack of ability to tap into the organizational-political network and gain power and influence can have a major detrimental affect on her ability to function effectively as a manager.

The view of organizations as political environments has been taken by a variety of organizational sociologists, most notably Dalton (1959), Crozier (1964), and Pettigrew (1973). More recently Schein (1977) has pointed out the neglect of this area by those in the field of organizational behavior and outlined a conceptual scheme as a basis for pursuing research in this area.

Although, as yet, there has been no integrated research in this area, descriptive research based on observations of managerial behavior and data does provide some assistance in documenting the existence of these power related behaviors. Martin and Simms (1956), for example, discuss the tactics practiced by most men whose success rests on the ability to control and direct the actions of others. Among the tactics they report are: taking counsel, forming alliances, maintaining maneuverability, promoting limited communication, compromising, negative timing, using self-dramatization, and exhibiting confidence. Strauss (1962), in an analysis of a purchasing agent's desire to expand his influence, outlines tactics of bureaucratic gamesmanship such as rule orientation, rule evasion, and

political. In a case analysis of a middle manager's tactics for power expansion, Izraeli (1975) observed such tactics as neutralizing potential opposition, making strategic replacements, committing the uncommitted, and forming a winning coalition. And finally DuBrin (1974) outlines such power acquisition strategies as: maintaining alliances with powerful people, embracing or demolishing, manipulating classified information, making a quick showing, bargaining, avoiding decisive action, and starting small.

Schein (1977) has categorized these behaviors into two types: work-related and personal power acquisition behaviors. The former are behaviors designed to achieve organizational objectives; whereas the latter are personal or political power acqusition behaviors related to personal aggrandizement such as through additional promotions, increased status, etc. Key in this discussion is the acknowledgement of the existence of and seeming need for the former. Strategies and tactics designed to increase one's power in the organization may be more related to managerial and organizational effectiveness than heretofore considered. For example, the powerless department head has less influence on the amount of money and promotional slots with which he or she can reward and motivate a subordinate; the powerless head of production has little bargaining power with the head of supplies so as to move out production faster; the powerless head of research has fewer resources with which to trade off so as to get his or her projects implemented, and so on. Work-related power acquisition behaviors may be necessary for effective managerial functioning within an organization.

Women and Work-Related Power Acquisition Behaviors

If such behaviors are as important to effective managerial functioning as the acquisition of technical and administrative skills, then there appears to be at least three implications of this with regard to women in managerial positions. First, it suggests that the current research examining the differential relationship between leadership style and satisfaction and productivity of subordinates as a function of sex of the leader may be less relevant than heretofore considered. Consideration of power and political power acquisition behavior suggests that there are a whole host of behaviors that a manager exhibits above and beyond the way in which he or she deals with his or her subordinates in the work group and these behaviors may be far more important to effective managerial functioning than leadership style. That supervisory behaviors outside of his work unit may be more related to subordinate satisfaction and productivity than "style" could be a partial answer to Terborg's question, "If stereotypes are as pervasive as some assume, then why are the results in this area so inconsistent?" (Note 2, p. 18).

A second implication with regard to the role of power and political behaviors and effective managerial functioning is with regard to women's ability to learn these strategies and tactics. Given that these power acquisition skills are not part of the formal training of managers, it seems that these are behaviors that are acquired on the job. Sex role stereotypical thinking with regard to women as managers may, however, limit their opportunities for

acqusition of these behaviors. For example, superiors with biased attitudes towards women may be less likely to openly discuss their strategies and tactics of operating within their organizations with their female subordinates. Or women may be more likely than men to be assigned to "at the desk" task-oriented aspects of a job, as opposed to the implementation aspects, thereby minimizing their awareness of the need for strategies and tactics in order to be effective at the total managerial job. Women, then, in comparison to men, may have fewer opportunities to acquire power acquisition behaviors or to be exposed to the necessity of exhibiting these behaviors as part of the managerial job.

Third, and perhaps most important, even if a woman is aware of the necessity to be strategic and to acquire power and influence, she may be excluded from what may be one of the most significant components of successful power acquisition--the development of informal/influence relationships. Among the key behaviors that emerge out of descriptive and observational work are those of forming alliances, coalition formation, influence trading, contact formation, etc. For example, Cyert and March (1964) describe the executive as a "political broker" operating in a system where the decisions on the allocation of resources are made by political coalitions, each with potential control over the system. Without such coalitions and alliances, a manager may be severely handicapped with regard to allocation of budget funds, promotional slots, specific project approvals, etc.

All of these liaisons and contacts revolve around an informal network of people built up over time through the development of relationships both on and off the job. Stereotypical thinking with regard to women, however, may foster both exclusion from these networks and/or make it more difficult for women to become active participants. For example, generalized negative attitudes towards women managers may prompt male managers to purposely not include women in the hopes that without these informal power bases they will be ineffective. Or stereotypical attitudes towards women may foster within males anxiety or discomfort about informal relationships such that they avoid any contact which is not directly job related. For example, a male may feel ill at ease lunching with a female peer or discussing "shop" after work hours. As such, the female is excluded from these informal opportunities to gain information about other functional areas, to learn of impending activities that may help or hinder her current projects or to influence other organizational members, outside of her formal communication network, as to her job-related needs.

The outcome of this exclusion may be the impairment of her ability to function effectively as a manager. To the extent that such alliances, trading of favors, and influence networks are important for getting the work done, a woman's probability for success as a manager would seem to be less than those of her male counterparts. And, to the extent that these power acquisition behaviors are rarely acknowledged by organizational members (Schein,

Note 3) her lack of effectiveness can then be attributed to her inability to perform the technical and administrative managerial skills. A woman may be subject to a very subtle but quite impactful form of discrimination by exclusion from these liaisons. Finally, to the extent that a woman is unaware of this aspect of managerial life she may accept this organizational view of her technical and administrative skill inadequacy and lower her own aspiration level.

Limited opportunities to acquire work-related power acquisition behaviors and exclusion from political/influence networks within organizations can, then, limit the performance effectiveness of the woman with the potential, as well as, in the long run, diminish her motivation to perform. As such, a prime research target should be the role of power acquisition behaviors in effective managerial functioning and the effect of sex and sex role stereotypical thinking on acquisition of and ability to use these behaviors.

In conclusion, both consideration of prior research and suggested directions for future research point to the need to focus on the performance side of the ability-performance relationship. As indicated earlier, there is already a reasonable body of research documenting the negative impact of sex role stereotyping on selection and perceptions of ability. Although not considered here, there is also a reasonable body of literature focusing on the impact of sex role stereotypical thinking on women's motivation to perform well. What is yet unknown is what <u>organizational</u> factors--be it distorted evaluations of her behaviors or performance or exclusion from learn-

ing and utilizing power acquisition behavior--serve to limit the performance of the woman in the organization.

Reference Notes

1. Hansen, P. Sex differences and supervision. Paper presented at the 82nd Annual Convention of the American Psychological Association, New Orleans, September, 1974.

2. Terborg, J. R. Integration of women into managerial positions: A research review. Paper presented at the 84th Annual Convention of the American Psychological Association, Washington, D.C., September, 1976.

3. Schein, V. E. Examining an illusion: The role of deceptive behavior in organizations. Paper presented at the Deception Conference, Department of Defense, New York, November, 1976.

References

Anastasi, A., & Foley, J. R., Jr. Differential psychology. New York: Macmillan, 1949.

Bigoness, W. J. Effect of applicants' sex, race, and performance on employer's performance rating: Some additional findings. Journal of Applied Psychology, 1976, 61, 80-84.

Brief, A. P., & Wallace, M. J. The impact of employee sex and performance on the allocation of organizational rewards. Journal of Psychology, 1976, 92, 25-34.

Cohen, S. L., & Bunker, K. A. Subtle effects of sex role stereotypes on recruiters' hiring decisions. Journal of Applied Psychology, 1975, 60, 566-572.

Crozier, M. The bureauractic phenomenon. London: Tavistock, 1964.

Cyert, R. M., & March, J. C. A behavioral theory of the firm. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Dalton, M. Men who manage. New York: Wiley, 1959.

DuBrin, A. J. Fundamentals of organizational behavior. Elmsford, N.J.: Pergamon Press, 1974.

Garland, H., & Price, K. H. Attitudes toward women in management and attributions for their success and failure in a managerial position. Journal of Applied Psychology, 1977, 62, 29-33.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. Race and sex as determinants of ratings by potential employers in a simulated work sampling task. Journal of Applied Psychology, 1974, 59, 705-711.

Izraeli, D. N. The middle manager and the tactics of power expansion: A case study. Sloan Management Review, 1975, 16, 57-70.

Maccoby, E. E. (Ed.). The development of sex differences. Stanford: Stanford University Press, 1966.

Martin, N. H., & Sims, J. H. Thinking ahead. Harvard Business Review, 1956, 34, 25-26.

Moses, J. L., & Boehm, V. R. Relationship of assessment center performance to management progress of women. Journal of Applied Psychology, 1975, 60, 527-529.

Pelz, D. C. Influence: A key to effective leadership in the first-line supervisor. Personnel, 1952, 9, 3-11.

Pettigrew, A. M. The politics of organizational decision making. London: Tavistock, 1973.

Petty, M. M., & Lee, G. K. Moderating effects of sex of supervisor and subordinate on relationships between supervisor behavior and subordinate satisfaction. Journal of Applied Psychology, 1975, 60, 624-628.

Rosen, B., & Jerdee, T. H. The influence of sex role stereotypes on evaluations of male and female supervisory behavior. Journal of Applied Psychology, 1973, 57, 44-48.

Rosen, B., & Jerdee, T. H. Effects of applicants' sex and difficulty of job on evaluations of candidates for managerial positions. Journal of Applied Psychology, 1974, 59, 511-512.

Rosenkrantz, P., Vogel, S., Bee, H., Broverman, I., & Broverman, D. Sex role stereotypes and self-concept in college students. Journal of Consulting and Clinical Psychology, 1968, 32, 287-295.

Schein, V. E. The relationship between sex role stereotypes and requisite management characteristics. Journal of Applied Psychology, 1973, 57, 95-100.

Schein, V. E. The relationship between sex role stereotypes and requisite management characteristics among female managers. Journal of Applied Psychology, 1975, 60, 340-344.

Schein, V. E. Individual power and political behaviors in organizations: An inadequately explored reality. Academy of Management Review, 1977, 2, 64-72.

Strauss, G. Tactics of lateral relationships: The purchasing agent. Administrative Science Quarterly, 1962, 7, 161-186.

Wylie, R. The self-concept. Lincoln: University of Nebraska Press, 1961.

The Interaction of Ability and Motivation in Performance: An Exploration of the Meaning of Moderators

Edwin A. Locke, Anthony J. Mento, and Bruce L. Katcher

University of Maryland, College Park

Abstract

It was hypothesized that one possible explanation of moderator effects is that they are due to different degrees of homogeneity with respect to a causal variable among different subgroups. This hypothesis was tested in a laboratory experiment in which performance was predicted from ability using motivation as the moderator. Ability was measured with a work sample and motivation was varied by assigning goals with different degrees of difficulty and specificity. It was found that ability predicted performance better in groups which were homogeneous with respect to motivation than in those which were motivationally heterogeneous. A moderated regression analysis showed that most of the differential validity was reducible to main effects, but significant interaction effects were found. One of them was caused by the fact that in some low motivation subjects the variance in performance is reduced, thus decreasing the slope of the regression line.

The Interaction of Ability and Motivation in Performance: An Exploration of the Meaning of Moderators[1]

Edwin A. Locke, Anthony J. Mento, and Bruce L. Katcher
University of Maryland, College Park

There has been persistent controversy over the issue of whether motivation and ability interact to produce performance (P). If ability (A) is defined as the capacity to perform (i.e., knowledge and skill), and motivation (M) is defined as the desire to perform (i.e., effort), then the interaction hypothesis is that $P = A \times M$. A second hypothesis, except for cases where A or M equal 0, is that A and M combine additively, thus $P = A + M$. A third alternative is to combine the first two, so that $P = A + M + (A \times M)$.

Since numerous studies have shown evidence for the main effects of both A and M, the real controversy is over the existence of the interaction term in the additive model.

The results of studies on this subject have been highly inconsistent. French (1958) found that high ability predicted performance only among high motivation subjects. In contrast, Fleishman (1958) and Lawler (1966) found that motivation predicted performance only for high ability individuals. On the other hand, Kipnis (1962) discovered that a test alleged to measure persistence was correlated with per-

[1]The authors gratefully acknowledge the helpful comments and suggestions of Dr. Philip Bobko on this paper.

formance only among low ability subjects. Finally, Howard (Note 1) and Locke (1965) found virtually no evidence for an interaction effect.

If one includes the expectancy (VIE) theory literature, taking V as the motivation variable and E as a correlate of ability, the evidence for a V x E interaction is equally uncertain (Korman, Greenhaus, & Badin, 1977).

The wider issue involved in this research clearly involves moderator variables (Zedeck, 1971). The findings of French (1958), Fleishman (1958), Kipnis (1962) and Lawler (1966) described above are all examples of differential validity. Furthermore, the form of the third equation above, if either motivation or ability is treated as a moderator variable, is of the same form as the moderated regression equation originally described by Saunders (1956; see also Zedeck, 1971).

Interestingly the state of our knowledge regarding the interaction of motivation and ability parallels the current status of moderator variable research. The findings have been highly inconsistent and no convincing explanation of the principles by which moderators operate has yet been offered.

It is generally assumed that subgrouping in moderator research is based on the principle of "homogeneity," with performance in homogeneous groups being more predictable than in non-homogeneous groups. However, the question: "Homogeneous with respect to what?" has not been answered.

One possibility is: Homogeneity with respect to measurement error in either the predictor or the criterion (including scoring errors, unreliability, criterion contamination, etc.). The attempt to use race as a moderator was based on the controversial hypothesis that conventional aptitude tests did not accurately measure the capacities of blacks (e.g., see Boehm, 1977; Katzell & Dyer, 1977). However, there have been successful moderator studies which involved subgrouping on the basis of what could be described as differences in measurement error. For example, Carroll and Nash (1972) found that a forced choice reference check was valid in predicting future job performance only for those applicants: who previously held jobs similar to the one applied for; whose previous supervisor knew the ratee for more than two months; and for whom sex, race and nationality differences with the rater did not produce bias.

It should be stressed that it is not accuracy as such that is the moderator here but rather homogeneity of accuracy. If the applicants in the above study who were deleted from the sample had all been given ratings which were equally biased in the same direction, their performance would have been as predictable as that of the applicants with accurate ratings. Typically, of course, the degree and direction of bias or error is unknown so that mismeasured groups are, in practice, heterogeneous in this respect.

A second, more complex, possibility is that moderator effects are caused by differences in homogeneity with respect to causal

factors. For example, it is logical to assume that ability will predict performance better in groups which are homogeneous with respect to motivation than in groups that are motivationally heterogeneous. Similarly, motivation should predict performance better in groups that are homogeneous in ability than in those that are heterogeneous in this respect.

Observe that the homogeneity hypothesis does not predict, as did a number of previous studies on the subject, that ability will predict performance better in high than in low motivation groups or that motivation will predict performance better in high than in low ability groups. The moderator effect proposed here is not based on the amount of the causal factor present (providing there is some amount present) but on the degree to which the amount is unequal among members of one group as compared with another group.

It should be noted that if our homogeneity hypothesis is fully correct, it would mean that any moderator effects (differential validity) obtained would actually be reducible to additive effects. In other words, the moderator effect obtained would be the result of unanalyzed or uncontrolled causes which, when identified, would produce only main effects. Such false moderator effects would be revealed by the lack of a significant interaction effect in a moderated regression analysis.

It is possible that many of the alleged moderator effects obtained in previous studies are actually the result of existing

but unmeasured main effects. For example, consider the finding that the academic performance of females is more predictable than that of males (Abelson, 1952). What could be the reason for this? One hypothesis would be that females are more homogeneous with respect to motivation (e.g., grade goals) than males, due probably to cultural reasons such as conformity to parents' wishes. If this were the case, then the moderating effect of sex would really be due to differential homogenizing with respect to another main effect, motivation.

(It should be added that a more careful examination of the homogeneity-with-respect-to-cause hypothesis shows that there are important exceptions and qualifications to it, but we will leave these to the discussion section.)

If our hypothesis is basically correct, it remains to explain the inconsistent results of the previous studies of ability-motivation interactions. Unfortunately these studies cannot be evaluated meaningfully, because the ability and/or motivation measures used were of unknown and often dubious validity (e.g., projective tests, self reports). A proper test of the homogeneity hypothesis would require measures of motivation and ability which were of unquestioned validity. Fortunately, such measures are available.

With respect to ability, the most valid measure of an individual's capacity to perform a specific task (in which new learning plays no significant role) is his performance over a short time span on a work sample, i.e., on the same task. Such a measure was used in the pre-

sent study.

With respect to motivation, it has been found in numerous studies (e.g., Locke, 1968) that the level of difficulty of an individual's performance goal is directly related to his actual performance level, given sufficient ability, and is therefore indicative of the degree of effort the individual is exerting.

There are at least two dimensions on which goals can vary. One is difficulty, as noted above. Combining subjects across difficulty levels should result in less homogeneity of motivation than would be present within any given difficulty level. A second attribute of goals is specificity. Individuals aiming for specific quantitative goals, providing they have feedback concerning their progress in relation to their goals (Locke, Cartledge, & Koeppel, 1968), should show more homogeneous levels of effort (at any given level of difficulty) than individuals given general or vague goals. Both methods of varying homogeneity of motivation were used in the present study.

The hypothesis of the present study was that moderator effects would occur when subjects were subgrouped according to their degree of homogeneity with respect to motivation. Specifically, it was hypothesized that ability would predict performance better in groups which were homogeneous with respect to motivation than in those which were heterogeneous in this respect.

Method

Subjects

The original subject pool consisted of 288 subjects drawn from the Introductory Psychology subject pool. Extra credit was given for participation. The subjects were run in groups. Eighteen sessions were run in all with the subjects in each session being given the same experimental treatment. Forty-six subjects were subsequently dropped from the analysis. A total of nine subjects in the three Specific goal conditions (described below) were dropped for performing incorrect calculations of their goals. The remaining 37 were dropped from the Specific-Medium and Specific-Easy groups for surpassing the assigned goal by an excessive amount, indicating a lack of full goal acceptance. Excessive was defined as surpassing the assigned goal by more than 30%. The number of subjects ultimately retained in each condition is shown below.

Task

The task was perceptual speed (Moran & Mefferd, 1959). The task requires the subject to indicate how many numbers in a row of 30 single digit numbers are the same as the first number in the row, which was circled.

Procedure and Measures

All subjects were first given a two-minute familiarization trial, followed by two two minute practice trials. Number correct on the

second practice trial was used as a basis for calculating the goals in the Specific goal conditions (described below), and was used as the ability measure in all subsequent analyses. The main work period in all groups was 20 minutes. Number correct in this period was used as the measure of performance. At the end of the 20-minute session all subjects were given a short questionnaire asking them to describe what goals they were trying for. All subjects indicated that they were trying to reach their goals. However, since the behavior of some indicated that they were trying to exceed their goals, and were thus not homogeneous with the rest of the subjects in that condition, they were removed from the analysis as described above.

Motivation Conditions

Six different motivation conditions were run; half the groups had specific goals with feedback while half had general goals without feedback. There were three levels of difficulty within each of the above two conditions. The six groups are described below.

Specific-Hard (N = 44). Subjects in this group were instructed to work at a pace equivalent to 100% of their pace on the second practice period. In order to know how much work this would involve, the second practice period work sheets were scored by the subject; the scores were then pro-rated so that the subject could mark in his main work booklet how far he should have gotten at the halfway point (which was announced) and by the end of the work period. For example, to get his overall goal, a subject in this group would

multiply his second practice trial score by 10; his halfway goal would be 1/2 of this amount.

Specific-Moderate (N = 27). The procedure was the same as in the previous group except that subjects were instructed to work at a pace representing 70% of their pace on the second practice period. They were instructed to mark their main work booklets accordingly.

Specific-Easy (N = 25). Subjects in this group followed the same procedure as above except that they were instructed to work at a pace which represented 30% of their pace on the second practice period. Again, they marked their work booklets accordingly.

General-Hard (N = 57). These subjects were told to "do their best" but were given no specific goals to aim for and no feedback.

General-Moderate (N = 52). These subjects were told to work at 70% of their second practice trial pace but were given no specific objective to reach and no feedback.

General-Easy (N = 37). These subjects were told to work at 30% of their second practice trial pace but were given no specific objective to reach and no feedback.

Results

To test the homogeneity hypothesis, two types of analyses were conducted. A differential validity analysis compared the ability-performance correlations among the different groups and combinations of groups. A moderated multiple regression analysis was employed to

identify main effects and interaction effects. The six treatment conditions were equal in initial ability ($F<1$). The ability-performance correlations for each experimental group as well as the performance means and standard deviations for these groups are shown in the upper left section of Table 1. The same information is shown for various combinations of groups in the right and lower portions of the table.

Differential Validity Analysis

Evidence relating to the differential validity hypothesis is shown in the last two rows of Table 1. The next to last row shows ability-performance correlations for the three specific goal groups combined, the three general goal groups combined and for all six groups combined, respectively. The last row shows the same correlations after standardizing within groups. The latter correlations are equivalent to the mean within-group correlations shown in the fifth row from the bottom.

The effect of homogeneity with respect to goal difficulty can be determined by comparing the combined raw vs. standardized ability-performance correlations (collapsing across specificity), since standardization homogenizes with respect to goal difficulty. The combined standardized correlation is .79 vs. a correlation of .49 based on raw scores (see bottom right section of Table 1). The difference between these correlations is significant at $p<.001$ by Hotelling's *t* test for non-independent r's.

Table 1

Performance Means and Standard Deviations, and

Correlations of Ability with Performance

Goal Difficulty		Goal Specificity: Specific	Goal Specificity: General	$\bar{r}$	Overall: $\bar{x}$	Overall: SD
Hard	r	.89[a]	.85	.87		
	$\bar{x}$	274	225		230	
	SD	33	37			35
Moderate	r	.94	.80	.88		
	$\bar{x}$	208	213		211	
	SD	43	36			38
Easy	r	.89	.43	.74		
	$\bar{x}$	76	160		126	
	SD	14	36			51
Overall	$\bar{r}$	.91	.73			
	$\bar{x}$	187	204			
	$\overline{SD}$	75	45			
Combined r					Combined r	
Raw Scores		.41	.65		.49	b
Standardized		.90	.73		.79	b
		c				

Notes: [a]All correlations are significant at $p<.01$ or better.

[b]Difference between non-independent r's significant at $p<.001$.

[c]Difference between independent r's significant at $p<.001$.

The effect of homogeneity with respect to specificity is revealed by comparing the overall standardized ability-performance correlations for the Specific vs. the General conditions (which are thus collapsed across goal difficulty). The correlation of .90 for the Specific groups is significantly greater than that of .73 for the General groups ($p < .001$).

We may conclude that ability predicts performance better in groups which are homogeneous with respect to goal difficulty and goal specificity than in groups which are heterogeneous in these respects.

It should be noted that collapsing across difficulty levels using raw scores shows the opposite pattern as compared to using standardized scores, with the overall correlation being higher for the General than the Specific groups (.65 vs. .41). This effect is due to the separation among the means being greater for the Specific groups; thus the combined Specific group is more heterogeneous in level of motivation than the combined General group, where the means were not initially separated to the same degree.

Moderated Regression Analysis

The analysis of main effects and interactions involved the use of multiple regression with effects coding for the treatment of categorical variables (Kerlinger & Pedhazur, 1973). Specifically, the Specific and General conditions were coded 1 and -1 respectively, while the Hard, Moderate, and Easy conditions were coded by creating

two vectors involving 1's, -1's and 0's. Various interactions were generated by vector multiplication. Using an hierarchical analysis approach was deemed appropriate since the variables could be entered in order of theoretical importance. The variable ordering was: ability, difficulty, specificity, ability x difficulty, specificity x difficulty, ability x difficulty x specificity plus ability x specificity.

A list of the order of variable entry, R^2, increment (increase in R^2 as a result of the entry of that variable), and significance is found in Table 2. The R^2 increment at each step in the analysis was tested by the F test for hierarchical variable entry found in Cohen and Cohen (1975).

The main effects of ability, difficulty, and specificity were significant ($p<.001$). A significant increment in variance was obtained when the ability x difficulty variable was entered. Thus, the regression slopes for difficulty collapsed across Goal specificity conditions were significantly different ($p<.001$). This means that goal difficulty moderated the linear relationship between ability and performance. It can be seen in Figure 1 that the slopes were less steep for the Easy Goal groups than for the remaining groups. There was also a significant specificity x difficulty interaction which can be observed in the pattern of means shown in Table 1. The remaining double interaction and the triple interaction were not significant.

Table 2

Moderated Regression Analysis

Variable	R^2	Increment	F
Ability	.239	.239	75.39***
Difficulty	.718	.479	202.11***
Specificity	.752	.034	32.38***
Ability x Difficulty	.770	.018	9.18***
Specificity x Difficulty	.878	.108	101.89***
Ability x Difficulty x Specificity plus Ability x Specificity	.880	.002	1.264 n

***p<.001

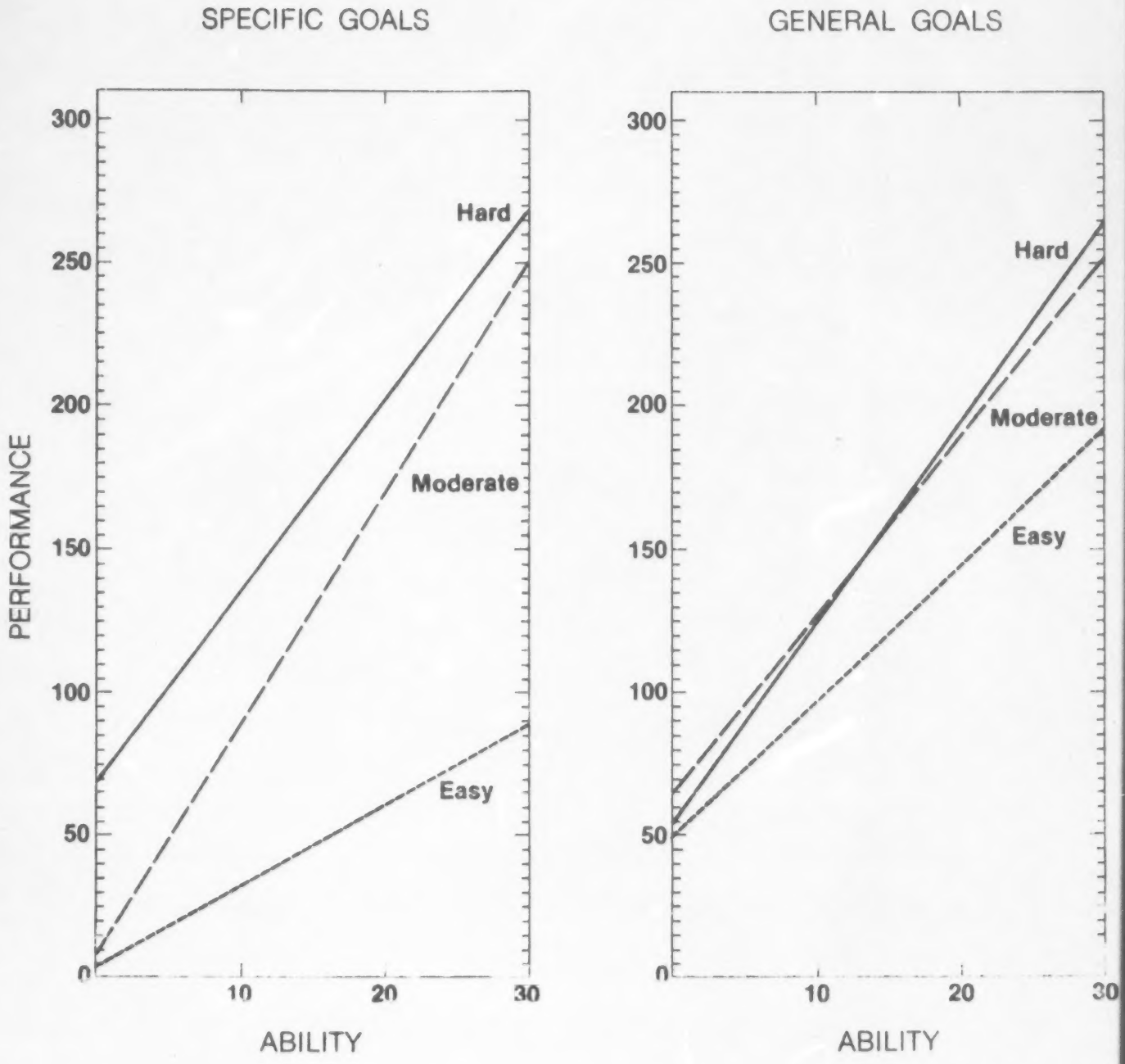


Figure 1. Regression lines for relationship of ability to performance for six experimental groups.

Discussion

We have shown convincing evidence that differential validity can result from variations among subgroups in degree of homogeneity with respect to a causal variable. It was shown that ability predicted performance better within groups that were homogeneous with respect to motivation than within groups which were heterogeneous in this respect.

Some previous moderator variable studies might be interpreted within this framework. For example, Raubenheimer and Tiffin (1971) found that ability tests predicted the clerical performance of employees who were married, conservative, stable and reserved better than the performance of clerks who were single, unstable, outgoing and bored by unchallenging jobs. It is probable that the latter group was more heterogeneous with respect to motivation since they would work hard if they were interested in their work but not if they were bored. Similarly, Hobert and Dunnette (1967) found that managers whose performance was underpredicted from ability and other tests were far more emotionally stable than those whose performance was overpredicted. These "unpredictable" managers were clearly more heterogeneous as a group (since it included both extremes of stability) than the managers who were predictable. Emotional stability itself may well have been an unanalyzed predictor in this study (Dunnette, 1972).

The latter point addresses an important implication of the present

results. It is that differential validity (including so-called qualitative moderators like sex) may often be largely reducible to linear effects once the causal variables are identified and measured. In effect, we are arguing that much of what we call differential validity may not involve true moderators, defined as a significant interaction term in a moderated regression equation, at all. There is only the appearance of moderators because the actual causes of the behavior in question have not been identified.

Most of the variance in the present study was reducible to the linear effects of ability, goal difficulty and goal specificity. The goal specificity effect was due entirely to the high performance of the General Easy subjects who performed way beyond the level of their assigned goal.

However, the findings of the present study indicate that the results were not entirely reducible to linear effects. Significant interaction terms were obtained. The specificity x difficulty interaction may be partly an artifact of the experimental situation. While the mean performance of the Specific Hard group was somewhat above that of the General Hard group (a common finding, see Locke, 1968), the performance of the Specific Easy group was substantially below that of the General Easy group. While neither of these groups were explicitly told to stop when they reached their assigned goal, the General Easy group was given no feedback regarding their performance in relation to the assigned goal. Thus there was no behavioral basis

for removing those members who might have deliberately exceeded their assigned rate from the analysis. In contrast, since the Specific Easy subjects had feedback, they could be and were removed from the analysis if they exceeded their goal by more than 30% (as noted earlier). More explicit experimental instructions might have reduced the performance of the General Easy subjects and thus eliminated much of the basis of the interaction.

The two way interaction between ability and difficulty was due to the generally lower slopes of the two Easy goal groups as compared to the Moderate and Hard group (Figure 1). However, the flatter slopes were due to different reasons in each case. The low slope of the General Easy group was due to the low ability-performance correlation (Table 1). The flat slope of the Specific Easy group was caused by the small variance in performance (slope = $\frac{s_y}{s_x} r$). The reason for, and mathematical necessity of, this effect is easily demonstrated. If each score in the Specific High group were multiplied by .3 (equivalent to 30% of maximum effort, which was the assigned goal of the Easy group), the SD of the performance scores in this group would be affected identically. The obtained SD of performance in the Specific Easy group was 42% of that of the Specific Hard group (Table 1).

(The s_y of the performance of the Specific Moderate group should have been 70% that of the Hard group but was not, due to the incomplete control of effort in that group.)

The ability-difficulty interaction, therefore, is at least partly a "real" one and is not in principle reducible to linear effects. Rather it is based on the effect of linear transformations of the performance scores of the less motivated subjects.

There are other examples in the literature of genuine interaction effects which are similar in principle to that described above. For example, it has been found that for employees with a stronger desire to grow on the job (those with "higher order need strength"), the correlation between degree of job enrichment and job satisfaction is higher than for subjects with a weaker desire to grow (Wanous, 1974). The presumed reason for this finding is that more important values have a greater impact on affect than less important values (Locke, 1976).

We must conclude, therefore, that the homogeneity with respect to cause hypothesis does not fully explain all alleged moderator effects based on causal variables.

There are other exceptions to the homogeneity principle as well. If, in a given group, a necessary condition for performance is entirely absent, ability (and/or motivation) will not predict performance in that group even though the group is homogeneous with respect to the missing cause. An illustrative example of this is a study by Erez (in press) which found that goal level predicted subsequent performance among subjects given adequate knowledge of their past performance but not among subjects given no feedback.

In conclusion it should be noted that there were minor flaws in this study which made the results somewhat less clear than they might have been; for example, insufficient control was attained over subjects exceeding their goals. Nevertheless the degree of control over extraneous variables and the degree of precision attained in measuring causal variables was far superior to that which is usually obtained in field studies. Additional laboratory studies of this type might throw further light on the nature of moderator effects, which, after 20 years of research, we still do not fully understand.

Reference Note

1. Howard, A. Intrinsic motivation and its determinants as factors enhancing the prediction of job performance from ability. Unpublished Ph.D. dissertation, University of Maryland, Department of Psychology, 1976.

References

Abelson, R. P. Sex differences in predictability of college grades. Educational and Psychological Measurement, 1952, 12, 638-644.

Boehm, V. R. Differential prediction: A methodological artifact? Journal of Applied Psychology, 1977, 62, 146-154.

Carroll, S. J., & Nash, A. N. Effectiveness of a forced-choice reference check. Personnel Administration, 1972, March-April, 42-46.

Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Lawrence Erlbaum, 1975.

Dunnette, M. D. Comments on Abrahams and Alf's "pratfalls in moderator research." Journal of Applied Psychology, 1972, 56, 252-256.

Erez, M. Feedback: A necessary condition for the goal setting-performance relationship. Journal of Applied Psychology, in press.

Fleishman, E. A. A relationship between incentive motivation and ability level in psychomotor performance. Journal of Experimental Psychology, 1958, 56, 78-81.

French, E. G. The interaction of achievement motivation and ability in problem-solving success. Journal of Abnormal and Social Psychology, 1958, 57, 306-309.

Hobert, R., & Dunnette, M. D. Development of moderator variables to enhance the prediction of managerial effectiveness. Journal of Applied Psychology, 1967, 61, 50-64.

Katzell, R. A., & Dyer, F. J. Differential validity revived. Journal of Applied Psychology, 1977, 62, 137-145.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, 1973.

Kipnis, D. A noncognitive correlate of performance among lower aptitude men. Journal of Applied Psychology, 1962, 46, 76-80.

Korman, A. K., Greenhaus, J. H., & Badin, I. J. Personnel attitudes and motivation. Annual Review of Psychology, 1977, 28, 175-196.

Lawler, E. E. Ability as a moderator of the relationship between job attitudes and job performance. Personnel Psychology, 1966, 19, 153-164.

Locke, E. A. Interaction of ability and motivation in performance. Perceptual and Motor Skills, 1965, 21, 719-725.

Locke, E. A. Toward a theory of task motivation and incentives. Organizational Behavior and Human Performance, 1968, 3, 157-189.

Locke, E. A. The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago. Rand McNally, 1976.

Locke, E. A., Cartledge, N., & Koeppel, J. Motivational effects of knowledge of results: A goal setting phenomenon? Psychological Bulletin, 1968, 70, 474-485.

Moran, L. J., & Mefferd, R. B. Repetitive psychometric measures. Psychological Reports, 1959, 5, 269-275.

Raubenheimer, I. W., & Tiffin, J. Personnel selection and the prediction of error. Journal of Applied Psychology, 1971, 55, 229-233.

Saunders, D. R. Moderator variables in prediction. Educational and Psychological Measurement, 1956, 16, 209-222.

Wanous, J. P. Individual differences and reactions to job characteristics. Journal of Applied Psychology, 1974, 59, 616-622.

Zedeck, S. Problems with the use of "moderator" variables. Psychological Bulletin, 1971, 76, 295-310.

Person-Situation Selection: A Review of Some Ability-Situation Interaction Research

Benjamin Schneider

Department of Psychology and Bureau of Business and Economic Research

University of Maryland, College Park

Abstract

A review of some experimental and field survey research regarding the effects of situation on ability-performance relationships is presented. The literature suggests that three factors, reward system, job characteristics, and leadership style/management philosophy ("climate") can moderate ability-performance relationships. These same three factors are known also to have main effects on performance and an hypothesis is presented suggesting that when reward system, job, and climate reward, support, and encourage the display of ability then validity for ability measures and performance levels will both be high.

Person-Situation Selection: A Review of Some Ability-Situation Interaction Research[1]

Benjamin Schneider

Department of Psychology and Bureau of Business and Economic Research

University of Maryland, College Park

In pursuit of increased validity, personnel selection researchers have concentrated their efforts on precision of measurement in the assessment of both predictors (test, interview, simulation) and criteria (turnover, sales style, quality of production). Indeed, to improve predictive validities, the use of multiple predictors and multiple or composite criteria is recommended (Dunnette, 1966; Guion, 1965; Schneider, 1976). Further, under the pressure of current federal legislation, differential validation studies are also required wherein the validity of a selection procedure is verified on, and for, different race and sex subgroups (cf. Guion, 1976) but this issue will not receive attention in the present paper (but see Bartlett, Dachler, Goldstein, & Schneider, Note 2; Howard, Note 3). Personnel psychologists have thus been concerned with refining techniques for selecting a best person from a number of people; their interest has been in individual differences.

---

[1]Portions of this paper appeared in Schneider (Note 1).

Organizational behaviorists have also been concerned with predicting behavior but they have been promoting ideas suggesting that the properties of organizations, not the personal attributes of people, are the important data in predicting and understanding behavior in the work setting. Theory Y, System 4, Consideration, Participation in Decision-Making; we are told that these are the organizational styles that result in increased effort and performance, decreased absenteeism and turnover, increased organizational commitment, decreased worker frustration, increased satisfaction, etc. (cf. Schein, 1970).

I, like other individual differences types, may be accused of overreacting to organizational behaviorists. Thus, one of the founders of the "O" branch of I-O psychology, McGregor (1960, p. 48) is rarely thought of as a personnel selection specialist yet two of his six Theory Y assumptions stated:

(a) "The capacity to exercise a relatively high degree of imagination, ingenuity, and creativity in the solution of organizational problems is widely, not narrowly, distributed in the population." (Italics mine)

(b) "Under the conditions of modern industrial life, the intellectual potentialities of the average human being are only partially utilized." (Italics mine)

Note here the emphasis in (a) on the distribution of ability (not everyone has equal ability) and, in (b), the constraints modern indus-

trial life puts on those differences (if one has averages, one has differences) being allowed to be expressed.

Unfortunately, however, the personnel selection and organizational behavior orientations to understanding and predicting behavior in the work setting have been following parallel rather than overlapping or integrated tracks (Porter, 1966); McGregor has had no impact on personnel selection practices. I shall argue here that there would be definite benefits for both selection and organizational researchers if an integrated view of the causes and correlates of employee behavior were developed. This integrated view would pay equal attention to individual differences (especially in ability) at the time of selection and to the kinds of situations organizations create for their human resources. The integrated view should result in: (1) Improved validity for selection assessment procedures; and, (2) A basis for understanding why increased levels of performance and satisfaction are found in organizations when certain organizational changes are made.

Person-Situation Research

Psychologists have talked person-situation research for many years and all good industrial-organizational psychologists know the Lewinian dictum $B = f(P, E)$. But scant research exists on the issue except for personnel selection researchers who long have known that ability only correlates with performance when the ability measured

is required by the job at which people work.

Recently there has been a resurgence of interest in person-situation research centered primarily on the personality-situation question generated by Mischel's (1968) extreme situationist position in his book, Personality and Assessment. More recently Endler and Magnusson (1976) have edited a volume on personality/situation interaction summarizing their own research and containing the influential papers of Bowers (1973) and Endler (1975).

A paper not included in the above volume but representative of the kind of research to be strived for in the work setting was accomplished by Andrews (1967). Briefly, Andrews administered T.A.T.s to the newly recruited management trainees of two Mexican corporations. One company was a dynamic, growth-oriented firm while the other was a relatively old, maintenance-oriented organization. All T.A.T.s were scored for nAch and nPow. Andrews found that 3 years later nAch was positively and significantly correlated with management progress in the dynamic company but negatively correlated in the older firm; exactly the reverse data were obtained for nPow.

Andrews' results revealed the potential for organizational attributes to have an effect on the relationship between a personal variable assessed prior to employment, nAch or nPow, and some subsequent on-the-job criterion of success. As Guion (1976, p. 798) noted with reference to selection in general, however,

The problem is that environmental factors influencing per-

> formance have not been considered very often in attempting predictions during the hiring process. . . . There are many potentially important situational variables, but only a few have been reported in studies relevant to selection. . . . [a]ny management practice which is suggested in the literature or folklore of management . . . is . . . appropriately considered as a possible predictor or moderator.

In the remainder of this review, some studies addressing the extent to which situational variables may affect ability-performance relationships are reviewed and an hypothesis for additional research is presented.

Situational Moderators of the Ability-Performance Relationship

Examples of both laboratory experimental and field survey studies exist that have examined ability-performance relationships as a function of situational differences (see Schneider, 1975, for other studies):

Laboratory Experimental Studies

Incentive and Task. Weinstein and Holzbach (1973) conducted a laboratory experiment with 72 male undergraduate students in which both reward system and task-flow interdependence were manipulated; the ability measure used was the Minnesota Clerical Tests. Although

their main interest was in the two manipulated variables, a differential validity analysis was also conducted.

This analysis revealed that: (1) Reward system had an impact on validity with people in an equal reward condition (same pay for all) being more predictable ($r = .46$, $p < .05$) than those in the differential reward condition ($r = .06$, n.s.); (2) Task-flow interdependence moderated validity with those in the low task flow interdependence condition more predictable ($r = .47$, $p < .05$) than those in the high task-flow independence condition ($r = .09$, n.s.); (3) The combined effects of task-flow interdependence and reward system had a quite dramatic effect on Minnesota Clerical validity as revealed in Table 1. The Table reveals that in an equal reward, low task-flow interdependence condition a correlation of .67 ($p < .01$) is obtained between the test and performance compared to a -.23 (n.s.) correlation in the equal reward, high task-flow interdependence.

Weinstein and Holzbach (1973, p. 299) concluded that:

> A limitation of the traditional personnel-differential approach to selection is its lack of accountability for mean differences in productivity across conditions. Typically, the researcher will develop a selection strategy for a job under a given set of conditions. The fact that a different set of conditions may yield a higher productivity is usually ignored. The experimentally oriented researcher is well aware of these mean differences, but typically ignores the

Table 1

Ability-Performance Correlations Under Different Reward and Task-Flow Conditions

	Task-Flow Interdependence		
Reward	Low	High	Combined
Differential	.24	.16	.08
Equal	.67*	-.23	.46*
Combined	.47*	.09	

*p .01

Note. The data in Table 1 are from Weinstein and Holzbach (1973).

gains in productivity which could be made through an assessment of differential ability. In the present study, the combined differential-experimental analysis accounted for twice as much variance as the experimental analysis and three times that of the differential analysis.

Studies on Incentive Systems. Dunnette (Note 4) has summarized research conducted at the University of Minnesota on the effects of reward systems on the ability-performance relationship. The three studies summarized were all attempts to understand the role of an incentive system and/or equity on performance and on the ability-performance relationship. All subjects were students working on real tasks for a real company at a temporary (6-day) clerical-type job (complete description of the experiments, excluding the ability-performance analyses, may be found in Pritchard, Dunnette, & Jorgenson, 1972 and Tornow, 1971).

The major conclusions from the Dunnette studies are paraphrased below:

1. Under equity conditions (compared to overreward and underreward conditions) ability is most consistently related to performance regardless of whether rewards are on an hourly or incentive basis.
2. Generally speaking, ability is reflected in performance more often under incentive reward conditions than under hourly reward conditions but these differences are not large ($\bar{r}$ = .75 vs. .65, respectively). These findings

might be thought to be somewhat different from those reported by Weinstein and Holzbach (1973) except for the fact that the latter authors only had an equity (equal pay) condition whereas Dunnette's studies had three conditions of equity. In fact, in Dunnette's equity condition the relationship between ability and performance is slightly higher for those under hourly pay ("true" equal, i.e., equity and hourly) than those under incentive pay.

3. Changes in reward systems have a depressive impact on ability-performance relationships. This effect is especially critical when moving from an incentive reward system to an hourly pay condition regardless of the equity condition.

4. In essentially every case average performance was superior under incentive than under hourly pay conditions.

Dunnette's results certainly fit Weinstein and Holzbach's (1973) caution about the effects of situation on levels of performance and ability-performance relationships. Dunnette stated (Note 4, p. 22):

> I conclude from these results that ability differences still are empirically the most important determiners of differences in job performance. The administrator's major purpose in trying to manipulate or alter incentive conditions becomes one of assuring the actual expression of those differences in the

form of job performance. . . .

and that the study suggests

> . . . that principles relating features of the task to be performed to the abilities required should . . . be based on variables that reflect the range of difficulty and the conditions of task performance.

Task Characteristics. Fleishman and his colleagues (Fleishman, 1957, 1972; Wheaton, Eisner, Mirabella, & Fleishman, 1976) have investigated the way in which task characteristics interact with ability in affecting performance. At a superficial level, Fleishman's work sounds like the typical personnel selection strategy - analyze the job, define the ability requirements and hire someone with the required attributes.

However, Fleishman has taken an approach that resulted in an investigation of ability-performance relationships as a function of the (experimentally manipulated) difficulty of the task. Thus in the most recent study (Wheaton et al., 1976), 127 male subjects took a battery of 24 tests, and performed an auditory signal detection task that was brought under experimental control (the difficulty manipulation) by manipulating signal duration and signal-to-noise ratio.

Results indicated that (Wheaton et al., 1976, p. 674):

> In general, the contributions of [the ability measure] to individual differences in performance increased as the

criterion task become more difficult.

A Study of Organizational Style. Frederiksen, Jensen, and Beaton (Note 5, 1972) conducted an innovative experiment designed to investigate the effects of organizational climate on, among other issues, predictor-criterion relationships. The study was conducted on 260 otherwise employed working male executives who participated in a two-day "Research Institute." Their task was the In-Basket Test (Frederiksen, Saunders, & Wand, 1957).

Each executive was randomly assigned to one of four experimental climates, Innovation, Rules, Global Supervision or Detailed Supervision. In addition the effects of compatible climates (Innovation/Global and Rules/Detailed) and incompatible climates (Innovation/Detailed and Rules/Global) were studied.

Through a series of complex analytic techniques, including three-mode factor analysis, Frederiksen et al. were able to show that relationships between ability and performance on particular kinds of In-Basket task factors varied as a function of the type of climate under which the executives worked.

Frederiksen et al. concluded with some implications for practice (Note 5, p. 357):

> The concept of differential predictability and the use of moderator variables recognize a greater degree of complexity in the prediction formula of the personnel psychologist. The notion is that the predictive value of one variable may be

> influenced by another variable, the moderator variable . . . . While most studies of moderator variables have employed measures of individual differences as the moderators, there is no reason why the moderator variable cannot be a situational variable.

Summary. Three different situational attributes, reward system (Dunnette, Note 4; Weinstein & Holzbach, 1973), task (Weinstein & Holzbach, 1973; Wheaton et al., 1976) and organization style (Frederiksen et al., 1972) were shown to have effects on ability-performance relationships when the situation was experimentally manipulated. As Guion (1976), noted, the range of potential situational moderators seems to encompass a number of organizational attributes thought to influence employee motivation and behavior.

Field Studies

Studies using similar situational parameters to those discussed above have been conducted in field settings by Lawler (1966) on reward systems, Howard (Note 3) as well as Bray, Campbell, and Grant (1974) on tasks, and Bray, Campbell, and Grant (1974) and Forehand (1968) on climate.

Reward System. Lawler's (1966) study of ability-reward system interaction used a number of potentially questionable methodological procedures. First, the reward system was indexed by employee views (N = 211) of the degree to which their pay was contingent on their

performance. Second, the ability "measure" was supervisory judgments of the incumbent's ability to do the job. Finally, the same supervisors served as raters of employee performance. However, although ability and performance ratings were significantly related ($r = .56$, $p < .01$), ability ratings were strongly related to education level ($r = .33$, $p < .01$) while performance was not, supporting the idea that the ability rating was based on somewhat different cues than was the performance rating.

In any case, by dichotomizing on reward system contingency views and the ability ratings the data shown in Table 2 were obtained. Analysis of variance results revealed main effects on performance ratings of both ability ($p < .01$) and contingency ($p < .05$) as well as an ability x reward contingency interaction ($p < .06$). As Lawler (1966, pp. 159-160) noted: ". . . [T]he performance scores supported the hypothesis that a multiplicative relationship exists between the contingency measure and ability."

Task Effects. In an extensive study of the role of "enriched" job characteristics as contributors to the prediction of performance from ability measures, Howard (Note 3) showed that task descriptions contribute significant (additive) variance to predictions.

Howard collected clerical ability data, job descriptions and supervisory ratings on 353 bank clerical personnel. Job descriptions were collected using the Job Diagnostic Survey (J.D.S.) developed by Hackman and Oldham (1975). The J.D.S. taps into five psychologically

Table 2

Performance Rankings for Different Ability Levels and Different Reward Systems

Ability Level	Performance Ranking	
	Low Contingency	High Contingency
High	50.3	54.7
Low	46.0	46.5

Note. Data in Table 2 are from Lawler (1966).

meaningful facets of tasks: skill variety, task identity, task significance, autonomy, and task feedback.

Howard (Note 3) regressed supervisory performance ratings on the ability data, a pooled J.D.S. score (average of the 5 J.D.S. facets), and their interaction. She found that both ability and job characteristics correlated with performance ($r = .18$, $p < .01$ and $r = .14$, $p < .01$, respectively), that a linear combination of the two yielded an $R = .25$, ($p < .01$) but that the interaction term (ability x job characteristics) yielded no <u>further</u> increment in predictability. However, although Howard does not report it directly it appears (p. 87) that the interaction term alone correlated $r = .23$ ($p < .01$) with performance.

Parenthetically, it should be noted that this issue of whether the multiplicative or additive formulation of person-situation effects on behavior is "better" has received considerable discussion (see Bowers, 1973; Endler, 1976; Howard, Note 3). Perhaps the most appropriate statement (at least the one we tend to agree with) was presented by Cummings and Schwab (1974, p. 46):

> Employee performance is seen as most directly a consequence of the employee's ability and his motivation to perform. . . . Certainly at the extremes of either ability or motivation some interaction must take place. Someone with no ability to complete a task cannot successfully perform no matter how highly motivated he may be to do so. Likewise, at least some modest amount of motivation is required, regardless of one's ability

to do a task, before we can expect successful performance. It is, however, much less clear that the notion of interaction contributes to the predictability of employee performance in applied settings where employees may be assumed to possess some minimal amounts of both ability and motivation. A simple additive approach will probably enable us to predict performance just about as well.

In any case, Howard (Note 3) clearly showed the merits of including the enriched nature of tasks in models attempting to understand the predictability of performance based on ability measures.

Climate. Under "climate" are included those studies that are concerned with more macro views of situational impacts on people (Schneider, 1975). While there are a number of studies that fit this definition (e.g., Schneider, Note 6; Vroom, 1960), two studies central to the ability-situation problem are reviewed, one by Forehand (1968) and one by Bray, Campbell, and Grant (1974).

Forehand (1968) reports research designed to explore the following questions (1968, p. 67): "Given a dimension on which individuals vary, what environmental conditions demand exercise of that quality, making the quality important for effective performance? And what environmental conditions constrain exercise of the quality, making the quality irrelevant to or incompatible with effective performance?"

Forehand adopted Frederiksen's et al. (1972) view of climate and divided 120 government executives into Group-Centered or Rules-Centered

work climates. Within each kind of climate he found the correlations reported in Table 3 between the various ability tests and peers' assessments of innovative behavior. The results reveal a strong effect of situation on ability-performance relationships.

Forehand (1968, pp. 75-76) concluded that studying person-situation interaction led to a clarification of the climate construct and then noted (p. 76):

> Such an analysis offers advantage to the study of individual differences (ID) as well as the study of climate. ID research has rested uncomfortably on the basis of the trait, and has unsuccessfully fielded such questions as: What is it? Where is it? Why doesn't it correlate with what it ought to correlate with? And why doesn't it correlate with the same variables from study to study? The analysis I have outlined suggests that the importance of an individual difference characteristic is precisely that it differentially predicts performance in different situations. Varying correlations between two measures are not an embarrassing evidence of error, but a dependent variable that we are interested in accounting for.

Bray, Campbell, and Grant (1974) report on the very comprehensive research conducted at A.T.&.T. in an attempt to understand the correlates and consequences, and the predictability of, managerial success and failure. Based on Assessment Center performance a prediction

Table 3

Ability-Performance Relationship Under Group-Centered and Rules-Centered Climates

	Peers' Assessment Ratings	
Test of Mental Alertness	Group-Centered (N = 60)	Rules-Centered (N = 60)
L Score	.43**	-.05
Q Score	.37**	.00
Total Score	.45**	-.03
Cognitive Factors		
Sensitivity to Problems	.36**	.01
Ideational Fluency	.30*	-.15
Penetration	.26*	-.03
Spontaneous Flexibility	.32**	.05
Semantic Redefinition	.31*	.01
Originality	.23	.13

*p .05

**p .01

Note. Data in Table 3 are from Forehand (1968).

was made of the likelihood of achieving the middle level of management within ten years. This prediction, then, was based on the recruit's own personal attributes. The predictions and management levels actually attained eight years later revealed that 64 percent of those predicted to reach middle management did so, while only 32 percent of those predicted to fail to reach middle management in fact reached it.

As Bray et al. (1974) note, however, the nature of a person's job may also have an effect on their subsequent success. In the Management Progress Study, jobs were characterized by the extent to which they contained more or less of each of eight dimensions: (1) Job stimulation and challenge, (2) Supervisory responsibilities, (3) Structured vs. unstructured assignments, (4) Objective stress of assignments, (5) Working alone vs. working in groups, (6) Morale of groups, (7) Supervision from bosses, and (8) Achievement models of bosses. Numbers 1, 2, 3, and 8 (the strongest correlates of success, $\bar{r} = .40$) were combined to form a Job Challenge index and the success rates of recruits were 59 percent, 31 percent, and 8 percent for those in high, moderate and low challenge jobs, respectively.

When the prediction based on personal attributes is cross-tabulated with the prediction based on job challenge the results reveal how both job and personal attributes combine in the prediction of performance achievement. These data are revealed in Table 4.

Bray et al. (1974, p. 76) commented that:

Table 4

Relationship of Assessment Prediction, Job Challenge, and Management Level Reached

Predicted to Reach	N	N Reach	% Reach
High Job Challenge	33	25	76
Moderate Job Challenge	22	12	55
Low Job Challenge	6	2	33
Total	61	39	64
Predicted to Fail to Reach			
High Job Challenge	18	11	61
Moderate Job Challenge	24	8	33
Low Job Challenge	20	1	5
Total	62	20	32

Note. From Formative Years in Business: A Long-Term AT&T Study of Managerial Lives, by D. W. Bray, R. J. Campbell, and D. L. Grant. New York: Wiley, 1974.

The extremes of the table reveal the powerful combined effect of job challenge and individual potential. Just over three-quarters of the more promising recruits who had had challenging jobs were in middle management eight years after employment, as compared with only one in 20 of the less promising recruits who were little challenged.

Summary. As with the earlier described experimental studies, reward system (Lawler, 1966), task (Howard, Note 3), and climate (Forehand, 1968; Bray et al., 1974) have been shown to contribute understanding to ability-performance relationships. Apparently each of these situational issues can have an influence on the extent to which people will do what they can do (Mace, 1935; Viteles, 1953).

Some Implications and an Hypothesis

Perhaps Forehand's (1968) view of the effects of the situation on the ability-performance relationship represents the most relevant basis for an hypothesis about how or why the effect occurs. Forehand alludes to constraints and facilitators of the display of the individual attribute of interest and the relevance of the situational issue for the performance of concern. Thus, assuming that the relevant ability or abilities for performance have been assessed, the critical issue seems to be to identify the facet or facets of the situation most likely to constrain or facilitate the display of that ability.

While we have made some modest progress in developing taxonomies of ability requirements in tasks (McCormick, 1974; Fleishman, 1972), almost no progress has been made in developing taxonomies of situations. This is not for a lack of trying. Frederiksen (1972) made the very useful distinction between taxonomies of attributes and taxonomies of situations, the former characterizing dimensions of situations, the latter characterizing types of situations and concluded that the most appropriate basis for defining types of situations was on the basis of the behavior they elicited. Frederiksen presented no such taxonomy but his suggestion to concentrate on the behavior of interest is instructive. The behavior of interest in the present context is, in fact, a relationship; a strong relationship between ability and performance.

The review of the ability-performance relationship presented above clearly suggests three issues, reward system, job characteristics, and organizational style/management philosophy that define a type of organization in which ability-performance relationships will be high. These are, of course, the same kinds of situational variables organizational behaviorists claim function as motivating conditions for obtaining increased levels of performance and job satisfaction. In terms of ability-performance relationships, then, these three conditions seem to function as constraints or facilitators of the display of ability while in terms of average levels of performance they function as stimuli or "triggers" to effort or performance. How may these

two sets of findings be integrated?

Figure 1 presents two scatter diagrams of joint distributions of ability and performance. In one case, portrayed in the bivariate distribution with the dotted line, the relationship between ability and performance is weak but positive. Also indicated with a dotted line is the average performance level for all people within the boundaries of the dotted line scattergram.

The second bivariate distribution in Figure 1, enclosed with a solid line, reveals a stronger ability-performance relationship. Note that this second distribution is narrower and extends higher on the performance dimension than the first distribution. That is, although the area within both distributions is similar, the distribution with the solid line represents a movement of people with higher levels of ability to higher levels of performance. Thus, low performing high ability people are now portrayed as high performers.

Note that this changes not only the strength of the relationship between ability and performance but also changes the average level of performance for the group. I hypothesize that this is precisely what happens when the kinds of organizational conditions discussed above exist in the work setting. Thus, what I propose is that under such conditions people are more likely to work up to their ability. Since work group performance is simply a function of how individuals perform, when those at the top of the ability distribution produce at a level that is commensurate with their potential, total work

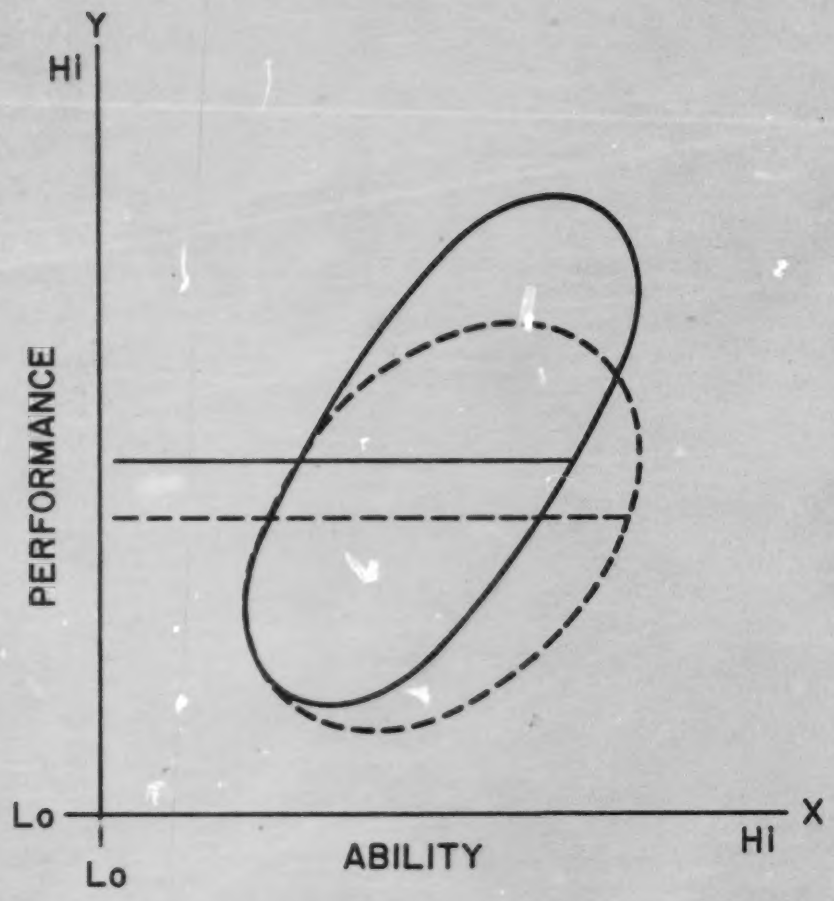


Fig. 1. Schematic for understanding increased average levels of performance when the ability-performance relationship is improved.

group performance must be generally higher than when high ability people perform below capacity. Those hypotheses are in keeping with results of studies discussed earlier by Dunnette (Note 4), Bray et al. (1974), and Weinstein and Holzbach (1973).

Of considerable interest is that increased levels of job satisfaction for employees should also follow. The idea that people will be more satisfied follows from the consistent finding that on challenging and enriching jobs, in more supportive organizations, and in organizations which reward people as individuals, employees tend to be more satisfied.[2]

One suspects that organizations have defined rigid rules of behavior for their employees so that they can gain control over individual differences; so they can accurately predict the behavior of aggregates of employees. It is paradoxical, but nevertheless apparently true, that just the opposite kind of orientation towards people, i.e., creating a climate supporting and rewarding the display of their abilities, will yield the same predictability of behavior with the added benefit of having higher average production and a more satisfied work force. Thus, although the potential to control behavior will have been taken away from management in a climate for individual differences, because accurate predictions will be possible, control would seem to be less necessary. It is precisely this lack of organization-

[2]Note that this does not mean that satisfaction and performance will be highly correlated because the reference here is for groups, not individuals.

ally imposed control that should yield the more satisfied work force.

Conclusion

Wise personnel selection decisions are at the foundation of an effective organizational behavior program in the work setting. People without requisite abilities cannot do their jobs effectively; attention only to their social/emotional state will not be helpful in producing a productive and satisfied work force.

On the other hand, appropriate organizational behavior practices can reward, support and encourage people to display the abilities they have. A good personnel selection system in such an organization will more likely be valid with concommintant higher levels of production and satisfaction.

Reference Notes

1. Schneider, B. Personnel selection and organizational behavior: An integrated view. Unpublished ONR Technical Report, Department of Psychology, University of Maryland, 1976.
2. Bartlett, C. J., Dachler, H. P., Goldstein, I. L., & Schneider, B. Enhancing the ability-performance relationship: A study of some psychological and contextual factors affecting total group and differential validity. Unpublished manuscript, University of Maryland, Department of Psychology, 1974.
3. Howard, A. Intrinsic motivation and its determinants as factors enhancing the prediction of job performance from ability. Unpublished ONR Technical Report, Department of Psychology, University of Maryland, 1976.
4. Dunnette, M. D. Performance equals ability and what? University of Minnesota, Department of Psychology, Technical Report No. 4009, 1973.
5. Frederiksen, N., Jensen, O., & Beaton, A. E. Organizational climate and administrative performance. Princeton, N.J.: Educational Testing Service, 1968.
6. Schneider, B. Organization type, organization success and the prediction of individual performance. Unpublished ONR Technical Report, Department of Psychology, University of Maryland, 1974.

References

Andrews, J. D. W. The achievement motive and advancement in two types of organizations. Journal of Personality and Social Psychology, 1967, 6, 163-168.

Bowers, K. S. Situationism in psychology: An analysis and critique. Psychological Review, 1973, 80, 307-336.

Bray, D. W., Campbell, R. J., & Grant, D. L. Formative years in business: A long-term A.T.&.T. study of managerial lives. New York: Wiley, 1974.

Cronbach, L. J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.

Cummings, L. L., & Schwab, D. P. Performance in organizations: Determinants and appraisal. Glenview, Ill.: Scott, Foresman, 1973.

Dunnette, M. D. Personnel selection and placement. Belmont, Calif.: Wadsworth, 1966.

Endler, N. S. The case for person-situation interactions. Canadian Psychological Review, 1975, 16, 12-21.

Endler, N. S., & Magnusson, D. (Eds.). Interactional psychology and personality. New York: Halstead, 1976.

Fleishman, E. A. Factor structure in relation to task difficulty in psychomotor performance. Educational and Psychological Measurement, 1957, 17, 522-532.

Fleishman, E. A. On the relation between abilities, learning, and human performance. American Psychologist, 1972, 27, 1017-1031.

Forehand, G. A. On the interaction of persons and organizations. In R. Tagiuri & G. Litwin (Eds.), Organizational climate: Explorations of a concept. Boston: Division of Research, Harvard Business School, 1968.

Frederiksen, N. Toward a taxonomy of situations. American Psychologist, 1972, 27, 114-123.

Frederiksen, N., Jensen, O., & Beaton, A. E. Prediction of organizational behavior. Elmsford, N.Y.: Pergamon, 1972.

Frederiksen, N., Saunders, D. R., & Wand, B. The in-basket test. Psychological Monographs, 1957, 71 (9, Whole No. 438).

Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

Guion, R. M. Recruiting, selection and job placement. In, M. D. Dunnette (Ed.), Handbook of industrial-organizational psychology. Chicago: Rand McNally, 1976.

Hackman, J. R., & Oldham, G. R. Development of the Job Diagnostic Survey. Journal of Applied Psychology, 1975, 60, 159-170.

Lawler, E. E., III. Ability as a moderator of the relationship between job attitudes and job performance. Personnel Psychology, 1966, 19, 153-164.

McCormick, E. J. Systematic job analysis. In D. Yoder & H. G. Heneman, Jr. (Eds.), Professional handbook of personnel management and industrial relations, Part 4. Washington, D.C.: Bureau of National Affairs, 1974.

McGregor, D. M. The human side of enterprise. New York: McGraw-Hill, 1960.

Mace, C. A. Incentives: Some experimental studies. Industrial Health Research Board (Report No. 72). London: H. M. Stationery Office, 1935. Cited in Vroom (1964).

Mischel, W. Personality and assessment. New York: Wiley, 1968.

Porter, L. W. Personnel management. Annual Review of Psychology, 1966, 17, 395-422.

Pritchard, R. D., Dunnette, M. D., & Jorgenson, D. O. Effects of perceptions of equity and inequity on worker performance and satisfaction. Journal of Applied Psychology, 1972, 56, 75-94.

Schein, E. H. Organizational psychology (rev. ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Schneider, B. Organizational climates: An essay. Personnel Psychology, 1975, 28, 447-479.

Schneider, B. Staffing organizations. Pacific Palisades, Calif.: Goodyear, 1976.

Tornow, W. W. The development and application of an input-outcome moderator test on the perception and reduction of inequity. Organizational Behavior and Human Performance, 1971, 6, 614-638.

Viteles, M. S. Motivation and morale in industry. New York: Norton, 1953.

Vroom, V. H. Some personality determinants of the effects of participation. Englewood Cliffs, N.J.: Prentice-Hall, 1960.

Weinstein, A. G., & Holzbach, R. L., Jr. Impact of individual differences, reward distribution, and task structure on productivity in a simulated work environment. Journal of Applied Psychology, 1973, 58, 296-301.

Wheaton, G. R., Eisner, E. J., Mirabella, A., & Fleishman, E. A. Ability requirements as a function of changes in the characteristics of an auditory signal identification task. Journal of Applied Psychology, 1976, 61, 663-676.

Summary and Implications of the Conference:

A Personal View

Benjamin Schneider

Department of Psychology and Bureau of Business and Economic Research

University of Maryland, College Park

Abstract

This note first presents a brief summary of the major points made by each of the contributions to the conference. The contributions are divided into four sets based on their focus: (1) methodological, (2) cautionary, (3) organization entry, and (4) organization process. Then implications of the presentations are presented, the two major ones being that (1) the chances of finding statistically significant algebraic interactions of ability and other variables (race, sex, situation, motivation) are poor; (2) in addition to more traditional methodological reasons (inadequate measurement), at a conceptual level algebraic interaction terms will not generally be found to be statistically significant in field studies because of naturally occurring interactions among people. The conclusion is reached that ability and work condition, in linear combination, should be used by personnel selection researchers to increase both the prediction of performance and the understanding of work behavior.

Summary and Implications of the Conference: A Personal View

Benjamin Schneider[1]
Department of Psychology and Bureau of Business and Economic Research
University of Maryland, College Park

The papers preceeding this brief summary and implications note were prepared for a conference at the University of Maryland sponsored by the Office of Naval Research. The goal of the conference was to encourage a diverse group of scholars to share their conceptual and methodological views about the establishment of ability-performance relationships in work settings. It was our feeling at Maryland that too much of the research in Industrial-Organizational (IO) Psychology remains in specific sub-disciplines of the field and that, particularly in the area of personnel selection, the focus of research has been overly narrow.

I say "particularly in personnel selection" because for the past decade or so researchers in this area of IO Psychology have been forced to concentrate their efforts almost exclusively on coping with socio-legal issues related to making actual selection decisions. During this same period of time, considerable progress was made in

---

[1]Discussion with my colleagues Phil Bobko, Pete Dachler and Ken Smith were helpful as I turned my taped comments from the conference into this note.

conceptualizing and researching the impact of organizational entry and organizational processes on work behavior, progress which seemed to offer the prospects of increased understanding of observed ability-performance relationships and increased predictability of performance. It was hoped that having people with diverse orientations together for a few days would broaden the participants' research horizons and help us be able to conduct our own ONR-sponsored research efforts at Maryland with knowledge of the most diverse methodological/conceptual schemes possible.

Although obviously presumptuous, I should like to share with you my view of the important points made by each paper and the implications of the presentations for future theory and research in understanding observed ability-performance relationships and in predicting performance.

Important Points

For me the papers divide into four sets: (1) the methodologically oriented, quantitatively focussed, papers of Bartlett, Bobko, Mosier and Hannan, and Schmidt and Hunter; (2) the "let's be careful about how we conceptualize these moderator variables" papers of Guion, and Owens; (3) the organizational entry paper by Wanous; and, (4) the organization process papers of Schein, Locke, Mento and Katcher, and Schneider.

Methodological Focus. Bartlett, et al. concentrated on the legal issue of race as a moderator of ability-performance relationships but, as was noted by the participants throughout the sessions, the data

analytic models they presented are applicable to the study of any interaction -- between ability and race, ability and sex, ability and access to power, or ability and goal-setting. Perhaps the most important general issue they raised was that an algebraic interaction can be considered statistically significant only when the beta weight attached to the interaction term is significant after the linear effects have been entered into the regression equation. It was clear from their presentation that only when a variable thought to moderate a relationship meets this criterion can it be called a moderator, and that few studies looking at race and sex as potential moderators achieve anything more than single group validity. Demonstrating single group validity does not indicate a moderator effect because the algebraic interaction will not be statistically significant.

Schmidt and Hunter presented a similar, but expanded argument. They noted that large-scale reviews of the literature on the validity of tests suggests the probabilities are high that one will find a significant relationship between an adequately chosen predictor and a carefully developed criterion when samples are large. That is, evidence accumulated over the years (prior probabilities) indicates we have been doing less poorly in accurately predicting job behavior than our typical small-sample studies lead us to believe.

Schmidt and Hunter proceeded to list a number of potential sources of error which, in any one study, prevent us from observing the "true" validity of a given ability measure. They further showed that when these sources of error are corrected for in a study, the potential

for finding a statistically significant algebraic interaction term between ability and some other personal (age, sex) or situational attribute (in their case, job type) is very small.

Bartlett, et al. argued, then, for the moderated multiple regression procedure (the "differential validity model") as the most appropriate analytic strategy for testing race or other variables as moderators. Schmidt and Hunter agreed with this admonition but added the insight that when the errors one normally makes in conducting validity studies are taken into account, the chances of finding statistically significant moderators are very small.

Let's-Be-Careful Focus. Owens' paper suggested the somewhat radical idea that an individual's group membership as determined by biodata is the best predictor of that individual's behavior. The logic of this paper was that moderator variables are irrelevant because all of the potential mathematically derived algebraic interactions in the prediction of behavior are taken into account by knowing a person's biodata group. That is, biodata group membership accounts for natural interactions making the generation or derivation of mathematical interactions redundant, at best. Owens marshalled a considerable amount of data to support his conclusions.

Guion made the useful observation that the word "ability" can be understood in a number of ways and that, from at least one perspective on that word, when performance on a job content sample is used as the ability measure in predicting job performance, "ability-performance relationship" is a redundancy. This line of thinking led him

to a consideration of the various meanings we attach to the phrase "content validity." Clearly he does not believe content validity is real in that it certainly has little to do with validity as we generally use the word in measurement. Indeed, Guion argued for us dropping the term from our vocabulary.

He suggested, in turn, that what we really do whenever we conduct a validity study is test a construct or set of constructs and that we must be far more explicit about the constructs we are testing than we have been in the past. Finally, he noted, following Owens' logic, that the many hypotheses we have about the nature of the personal attributes required for adequate job performance may be assessed and then profiled and that people might be selected for particular jobs based on their group membership or that group membership itself could be treated as a potential moderator.

Thus, the Owens and Guion papers both proposed alternatives to traditional personnel selection research strategies. Owens proposed that we employ the individual as the unit of analysis only in defining groups and then use group membership as the predictor. Guion suggested that our dust-bowl empiricism of the past requires modification, one modification being the conceptualization of job content samples for prediction purposes to be part of a more comprehensive process of construct validation.

Organization Entry Focus. Wanous' contribution was enlightening because his review of the literature on realistic job previews (RJP's) revealed no attempts to concurrently study ability-performance

relationships and the impact of RJP's on that relationship or the contribution of RJP's to the prediction of performance. Thus, the literature concerned with the organizational entry process remains unconnected to the prediction of individual behavior. Wanous' conceptualization of the issue suggests that one need only believe in an interaction between ability and motivation in order to conduct such studies.

Another aspect of Wanous' presentation that was of interest was the paucity of research available on the entry part of the selection process - it is as if selection researchers have forgotten that whole people go through whole processes and that those processes may have an impact on the validity and utility of their prediction devices or serve as an additional predictor of behavior.

Organization Process Focus. Schein's paper noted the debilitating effects of sex-role stereotyping on women expressing their abilities at work. One could make the argument that Schein's concerns about the debilitating effects of stereotyping are as applicable to the entry process (the selection, placement, and training opportunities organizations provide for women) as they are to organization process but I think her emphasis on power acquisition, and the power process in organizations, makes her view particularly salient to the realities of organizational life after the also-important entry process.

For me the major point made by Schein was her discussion of the ways in which women are systematically excluded from positions of

power with the concommitant lack of opportunity to express the full range of their abilities. Simply put, stereotyping women into a narrow range of jobs must, of course, lead to the blockage of women expressing the full range of individual differences that exist. This blockage keeps the level of performance of women artificially low.

Locke, Mento and Katcher also addressed the problem of enforced homogeneity noting that when people are homogeneous with respect to some causal psychological variable (e.g., motivation) their individual differences in performance will be more predictable from ability measures. In essence, their point was that when one controls for extraneous variables (like motivation) in trying to predict performance from ability measures, one's chances of being accurate are improved.

The experiment presented on goal-setting tended to support the hypothesis. Perhaps of greater significance, the data analyses presented in the paper can be interpreted as making a very important distinction between the prediction of individual differences in performance and the prediction of average performance level. Thus, the Locke, et al. paper clearly showed equal predictability of individual differences in performance in the three goal-setting conditions. However, because the average level of performance in the three goal-setting conditions was so different, prediction of performance level was enhanced when the ability data were combined with the goal-setting condition data. Indeed, both main effect (ability + goal-

setting) and interaction effects (ability + goal-setting + (ability x goal-setting)) were significant.

Schneider's paper examined some research that explored the contributions of ability and situation variables to the prediction of performance. In retrospect it must be noted that he tended to be quite casual in his use of the word "interaction" as in his title which refers to "ability-situation interaction research." In fact, of the field studies reviewed by Schneider just about all failed to test for the significance of an algebraic interaction term and most of the studies reviewed were not clear regarding the distinction I made earlier in discussing the Locke, et al., paper between predicting individual differences in performance and the prediction of performance level. I will return to this issue below.

An important outcome of Schneider's review was the paucity of studies that have apparently attempted to study the utility of using both ability and situation indices in the prediction of performance. This is surprising because of the constant admonition to selection researchers to "know thy situation" and the assumption made by organizational psychologists that people in the organizations they study possess the abilities to do their jobs.

Implications of the Presentations

I see two major implications of the presentations: (1) The chances of finding statistically significant algebraic interactions of ability and other variables (race, sex, situation, motivation) are poor. However, the fact that organizational processes such as new

employee entry, sex-role stereotyping, and more traditional variables of concern to organizational psychologists like goal-setting, reward system, supervision, and job enrichment contribute, both theoretically and empirically, to the prediction of performance from ability measures by having significant, independent, linear effects on performance level is liberating; (2) Thinking is required about why algebraic interaction terms are not generally found to be statistically significant and on defining conditions under which significant algebraic interaction terms may be observed. It is proposed that this thinking begin with a distinction being made between the effects of the normal interactions of people at work on the creation of work conditions and the phrase "statistically significant algebraic interaction term."

Finding Statistically Significant Algebraic Interaction Terms. Bartlett, et al. set forth the requirements for designating a variable as a moderator and showed that the largest proportion of studies that examined race and sex as potential moderators failed to meet those standards. Schmidt and Hunter's review suggested that statistically significant algebraic interaction terms are rare even when job variables are employed as potential moderators. Even in the laboratory experiment reported on by Locke, et al., the significant algebraic interaction term added precious little variance to the predictions of performance possible using the content sample (pretrial performance) index of ability and goal-setting condition, in linear combination. Owens indicated that moderators are unnecessary

when one knows a person's group membership. Are there statistically significant algebraic interaction terms to be found in conducting research on the prediction of performance from ability? I conclude the chances of finding them are small. Are variables like motivation, race, sex, and situation important contributors to the prediction of performance from ability measures? I conclude the answer is yes.

In retrospect, the problem is that whenever we think of predicting performance by combining the effects of ability and some other variable, we think of a multiplicative combination implying not only effects on intercepts ("main effects") but slopes ("interactions") as well. Much of the research reviewed by Schneider, for example, made this assumption implicitly. Theoretical ideas and data presented by Locke, et al., indicate, however, that a major impact of motivation (or situation) may be to simply affect intercepts, not slopes. The observed effects of motivational condition in their research were: (1) similar correlations between ability and performance within each condition; (2) differences in performance level between motivation conditions; (3) a depressed correlation between ability and performance <u>across</u> conditions; (4) main effects for both ability and condition in predicting performance; and, (5) a relatively high multiple R when ability and motivation condition are linearly combined to predict performance.

These conclusions may require elaboration. Regarding (1), Locke et al. showed that the ability performance correlations in each motivation condition were essentially the same. However, with

respect to (2), their data revealed clear differences in average performance levels for subjects in the different motivation conditions. The effect of (1) and (2) would be a depressed correlation between ability and performance if the relationship were examined for all subjects, regardless of motivation condition; thus conclusion (3) is stated. Conclusions (4) and (5) follow because ability and motivation condition are both related to performance but they are independent of each other; each, then contributes unique variance to the prediction of performance.

The importance of the finding that ability and motivation condition combine linearly to predict performance is that once the linear, rather than the multiplicative combination is accepted as the more appropriate formulation, it becomes easier to conceptualize, and do research on, ways of enhancing the prediction of performance from ability measures. Researchers become liberated from the worry about creating algebraic interaction terms in the absence of the required ratio scales; no longer must they debate over whether the algebraic interaction term or the linear effects must enter the regression equation first; and, no longer can personnel and organizational psychologists go their own merry ways assuming that they can discount the effects each other is able to show. In fact, each group of researchers is probably addressing a different part of the same problem, the selection researcher the problem of "who will perform how" and the organizational psychologist "how high will they perform." The selection researcher, by failing to account for motivational condition

(a criterion contaminant for Schmidt and Hunter) in the analysis of ability measure validity data, has trouble revealing the "true" validity of his measure. The organizational psychologist, on the other hand, by discounting individual differences in ability has trouble understanding why his organizational interventions "work" to increase performance in one job or company but fail in another.

In Wanous' conceptual scheme, and in the ideas presented by Schein, RJP's and power acquisition, respectively, become main effects on performance level. In combination with appropriate ability measures, such main effects should yield more accurate predictions of performance and performance level than the effects of either one alone. Similar results would be expected from job enrichment, reward system, organizational climates, and other performance-relevant organization variables as outlined by Schneider.

Why Don't the Algebraic Interaction Terms Reach Statistical Significance?

It is imporant to speculate about why the algebraic interactions between ability and motivation or situation fail to exist. Thus, while Bartlett, et al., and Schmidt and Hunter give some information about why race and sex fail to serve as true moderators, why am I so despondent about finding statistically significant mathematical interaction terms between ability and some other psychological variable like motivation or incentive system or job enrichment when performance is being predicted?

First, one expects statistically significant algebraic interaction

terms when one or both of the variables employed as independent variables or predictors have extreme scores. In most work settings extreme scores on variables like ability and motivation simply do not exist. Career and vocational choice literatures suggest, for example, that people do, indeed, seek work settings they more or less fit.

Second, creating algebraic interaction terms by multiplying variable scores requires a level of measurement not usually found in behavioral science research. In any case, even when some algebraic interaction terms are observed to be statistically significant, simple transformations of the data reduces them to linear effects.

Third, and perhaps most importantly, algebraic interaction terms may be conceptually redundant in research already employing a work condition variable as a predictor of performance. By this I mean that because the natural interaction of people in work settings creates the very work conditions being used as a predictor, the creation of algebraic interaction terms over and above a linear combination of ability and situation would be a redundancy.

Why do Locke, et al. and other laboratory researchers find statistically significant algebraic interaction terms? I suspect that in artificially created, short-term, laboratory studies where naturally occurring interactions are usually not permitted, that algebraic interaction terms may be significant. Perhaps a similar kind of thinking helps explain the few interactions of ability and job type reported by Schmidt and Hunter. That is, it is possible that the jobs

in question either require very low levels of interaction and/or the data were collected on newly formed groups (and thus laboratory-like conditions).

Whether or not the reader accepts the conclusions I derived from the papers, it seems clear to me that the conference accurately indicated the view that more issues than creating better tests of ability require attention. In particular, the conference participants noted that the influence of what happens in work settings on our predictions of performance from ability (or other selection measures of individual differences) has been sorely neglected. Perhaps when personnel selection researchers realize that (1) personnel selection procedures should contribute to the prediction of levels of performance, not just the establishment of a correlation between a test and a criterion; (2) a great deal of the interaction in work settings can be quantified and is predictive of average levels of performance; and (3) linear combinations of variables, not algebraic interactions, are sufficient to capture two effects on performance, we will begin to conduct the kind of holistic research that will yield both increased prediction and understanding of work performance.

Distribution List

Navy

4 Dr. Marshall J. Farr, Director
Personnel & Training Research Programs
Office of Naval Research (Code 458)
Arlington, VA 22217

1 ONR Branch Office
495 Summer Street
Boston, MA 02210
Attn: Dr. James Lester

1 ONR Branch Office
1030 East Green Street
Pasadena, CA 91101
Attn: Dr. Eugene Gloye

1 ONR Branch Office
536 S. Clark Street
Chicago, IL 60605
Attn: Dr. Charles E. Davis

1 Dr. M. A. Bertin, Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco 96503

1 Office of Naval Research
Code 200
Arlington, VA 22217

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Director, Human Resource Management
Naval Amphibious School
Naval Amphibious Base, Little Creek,
Norfolk, VA 23521

1 LCDR Charles J. Theisen, Jr., MSC, USN
4024
Naval Air Development Center
Warminster, PA 18974

1 CDR Paul D. Nelson, MSC, USN
Naval Medical R&D Command (Code 44)
National Naval Medical Center
Bethesda, MD 20014

1 Commanding Officer
Naval Health Research Center
San Diego, CA 92152
Attn: Library

1 Chairman, Leadership & Law Dept.
Div. of Professional Development
U.S. Naval Academy
Annapolis, MD 21402

1 Scientific Advisor to the Chief of
Naval Personnel (Pers Or)
Naval Bureau of Personnel
Room 4410, Arlington Annex
Washington, DC 20370

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Mr. Maurice Callahan
NODAC (Code 2)
Dept. of the Navy
Bldg. 2, Washington Navy Yard
(Anacostia)
Washington, DC 20374

1 Office of Civilian Personnel
Code 263
Washington, DC 20390

1 Superintendent (Code 1424)
Naval Postgraduate School
Monterey, CA 93940

1 Dr. H. M. West III
Deputy ADCNO for Civilian Planning
and Programming (Acting)
Room 2625, Arlington Annex
Washington, DC 20370

1 Mr. George N. Graine
Naval Sea Systems Command
SEA 047C12
Washington, DC 20362

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
Attn: Dr. Norman J. Kerr

1 Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508
Attn: Dr. William L. Maloy

1 Dr. Alfred F. Smode, Director
Training Analysis & Evaluation Group
Department of the Navy
Orlando, FL 32813

1 Chief of Naval Education and
Training Support (01A)
Pensacola, FL 32509

1 Naval Undersea Center
Code 303
San Diego, CA 92132
Attn: W. Gary Thomson

1 Navy Personnel R&D Center
Code 01
San Diego, CA 92152

5 A. A. Sjoholm, Head, Technical
Support
Navy Personnel R&D Center
Code 201
San Diego, CA 92152

2 Navy Personnel R&D Center
Code 310
San Diego, CA 92152
Attn: Dr. Martin F. Wiskoff

1 Dr. Robert Morrison
Navy Personnel R&D Center
Code 301
San Diego, CA 92152

1 Navy Personnel R&D Center
San Diego, CA 92152
Attn: Library

1 Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego, CA 92152

Army

1 Technical Director
U.S. Army Research Institute for
the Behavioral & Social
Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Armed Forces Staff College
Norfolk, VA 23511
Attn: Library

1 Commandant
U.S. Army Infantry School
Fort Benning, GA 31905
Attn: ATSH-I-V-IT

1 Commandant
U.S. Army Institute of Admin-
istration
Attn: EA
Fort Benjamin Harrison, IN
46216

1 Dr. Ralph Dusek
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Ralph Canter
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Milton S. Katz, Chief
Individual Training & Perfor-
mance Evaluation Technical
Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Research Branch
AFMPC/DPMYP
Randolph AFB, TX 78148

1 Dr. Marty Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Dr. Alfred R. Fregly
AFOSR/NL, Building 410
Bolling AFB, DC 20332

1 Air Force Human Resources Lab
AFHRL/PED
Brooks AFB, TX 78235

1 Major Wayne S. Sellman
Chief, Personnel Testing
AFMPL, DPMYO
Randolph AFB, TX 78148

1 Air University Library
AUL/LSE 76-443
Maxwell AFB, AL 36112

Marine Corps

1 Director, Office of Manpower Utilization
HQ, Marine Corps (Code MPU)
BCB, Building 2009
Quantico, VA 22134

1 Dr. A. L. Slafkosky
Scientific Advisor (Code RD-1)
HQ, U.S. Marine Corps
Washington, DC 20380

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-P-1/62)
U.S. Coast Guard Headquarters
Washington, DC 20590

Other DoD

1 Dr. Harold F. O'Neil, Jr.
Advanced Research Projects Agency
Cybernetics Technology, Room 623
1400 Wilson Boulevard
Arlington, VA 22209

1 Mr. Frederick W. Suffa
Chief, Recruiting & Retention Evaluation
Office of the Assistant Securetary of Defense, M&RA
Room 3D970, Pentagon
Washington, DC 20301

12 Defense Documentation Center
Cameron Station, Building 5
Alexandria, VA 22314
Attn: TC

1 Military Assistant for Human Resources
Office of the Director of Defense Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 Director, Management Information Systems Office
OSD, M&RA
Room 3B917, The Pentagon
Washington, DC 20301

Other Government

1 Dr. Lorraine D. Eyde
Personnel R&D Center
U.S. Civil Service Commission
1900 E Street, N.W.
Washington, DC 20415

1 Dr. William Gorham, Director
Personnel R&D Center
U.S. Civil Service Commission
1900 E Street, N.W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel R&D Center
U.S. Civil Service Commission
1900 E Street, N.W.
Washington, DC 20415

1 U.S. Civil Service Commission
Federal Office Building
Chicago Regional Staff Division
Regional Psychologist
230 S. Dearborn Street
Chicago, IL 60604
Attn: C. S. Winiewicz

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

1 Robert W. Stump
National Institute of Education
Washington, DC 20208

Miscellaneous

1 Mr. Samuel Ball
Educational Testing Service
Princeton, NJ 08540

1 Dr. Gerald V. Barrett
University of Akron
Department of Psychology
Akron, OH 44325

1 Dr. Bernard M. Bass
University of Rochester
Graduate School of Management
Rochester, NY 14627

1 Century Research Corporation
4113 Lee Highway
Arlington, VA 22207

1 Dr. Kenneth E. Clark
College of Arts and Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
Department of Psychology
University of Southern California
University Park
Los Angeles, CA 90007

1 Dr. John J. Collins
Essex Corporation
201 N Fairfax St.
Alexandria, VA 22314

1 Dr. Joseph E. Campoux
School of Business & Administration
University of New Mexico
Albuquerque, NM 87131

1 Prof. W. W. Cooper
Graduate School of Business Administration
Harvard University
Boston, MA 02163

1 Dr. Rene V. Dawis
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

1 Dr. Ruth Day
Department of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 Dr. Robert Dubin
University of California
Graduate School of Administration
Irvine, CA 92664

1 Dr. John D. Carroll
Psychometric Lab
Davie Hall 013A
University of North Carolina
Chapel Hill, NC 27514

1 Dr. Marvin D. Dunnette
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Major I. N. Evonic
Canadian Forces Personnel Applied Research Unit
1107 Avenue Road
Toronto, Ontario, CANADA

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Edwin A. Fleishman
Advanced Research Resources Organization
8555 Sixteenth Street
Silver Spring, MD 20910

1 Dr. John R. Frederiksen
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

1 Dr. Robert Glaser, Co-Director
University of Pittsburgh
3939 O'Hara Stteet
Pittsburgh, PA 15213

1 Dr. Gloria L. Grace
System Development Corporation
2500 Colorado Avenue
Santa Monica, CA 90406

1 Dr. Richard S. Hatch
Decision Systems Association, Inc.
5640 Nicholson Lane
Rockville, MD 20852

1 Dr. M. D Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean, VA 22101

1 HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921
Attn: Library

1 HumRRO/Columbus Office
Suite 23, 2601 Cross Country Drive
Columbus, GA 31906

1 HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921
Attn: Dr. Robert Vineberg

1 Dr. Lawrence B. Johnson
Lawrence Johnson & Associates, Inc.
Suite 502
2001 S Street, N.W.
Washington, DC 20009

1 Dr. Steven W. Keele
Department of Psychology
University of Oregon
Eugene, OR 97403

1 Mr. W. E. Lassiter
Data Solutions Corp.
Suite 211
6849 Old Dominion Drive
McLean, VA 22101

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Mr. Brian McNally
Educational Testing Service
Princeton, NJ 08540

1 Dr. Ernest J. McCormick
Department of Psychological Sciences
Purdue University
Lafayette, IN 47907

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Corton Drive
Santa Barbara Research Park
Goleta, CA 93017

1 Mr. Edmond Marks
304 Grange Bldg.
Pennsylvania State University
University Park, PA 16802

1 Dr. Leo Munday
Houghton Mifflin Co.
P.O. Box 1970
Iowa City, IA 52240

1 Richard T. Mowday
College of Business Administration
University of Oregon
Eugene, OR 97403

1 Mr. Luigi Petrullo
2431 N Edgewood Street
Arlington, VA 22207

1 Dr. Steven M. Pine
n 660 Elliott Hall
University of Minnesota
75 East River Road
Minneapolis, MN 55455

1 Dr. Lyman W. Porter, Dean
Graduate School of Administration
University of California
Irvine, CA 92717

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu, CA 90265

1 R. Dir. M. Rauch
P II 4
Bundesministerium der Verteidigung
Postfach 161
53 Bonn 1 GERMANY

1 Dr. Joseph W. Rigney
University of So. alifornia
Behavioral Technology Laboratories
3717 South Grand
Los Angeles, CA 90007

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. N.W.
Washington, DC 20007

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Lyle Schoenfeldt
School of Management
Rensselaer Polytechnic Institute
Troy, NY 12181

1 Dr. Mark D. Reckase
Educational Psychology Department
University of Missouri-Columbia
12 Hill Hall
Columbia, MO 65201

1 Dr. Richard Snow
Stanford University
School of Education
Stanford, CA 94305

1 Dr. C. Harold Stone
1428 Virginia Avenue
Glendale, CA 91202

1 Dr. David J. Weiss
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

1 Dr. Earl Hunt
Department of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Thomas G. Sticht
Assoc. Director, Basic Skills
National Institue of Education
1200 19th Street N.W.
Washington, DC 20208

1 Prof. Fumiko Samejima
Department of Psychology
Austin Peay Hall 304C
University of Tennessee
Knoxville, TN 37916

1 Dr. John Wanous
Department of Management
Michigan State University
East Lansing, MI 48823

1 Dr. Frank Pratzner
The Center for Vocational Education
Ohio State University
1960 Kenny Road
Columbus, OH 43210

1 Dr. Meredith Crawford
5605 Montgomery Street
Chevy Chase, MD 20015